NPS55-83-029PR

NAVAL POSTGRADUATE SCHOOL

Monterey, California

ANALYSIS OF POTENTIAL OUTLIER
AND MISSING DATA POINTS
IN NUWES DATA

by

J. B. Tysver

October 1983

Prepared for: Commanding Officer
Naval Undersea Warfare Engineering Station
Keyport, WA 98345

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral J. J. Ekelund
Superintendent

David A. Schrady
Provost

The work herein was supported in part by funds provided by the Naval Undersea Warfare Engineering Station, Keyport, Washington.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER NPS55-83-029PR	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ANALYSIS OF POTENTIAL OUTLIER AND MISSING DATA POINTS IN NUWES DATA		5. TYPE OF REPORT & PERIOD COVERED Project Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) J. Bryce Tysver		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93943		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Commanding Officer Naval Undersea Warfare Engineering Station Keyport, WA 98345		12. REPORT DATE October 1983 13. NUMBER OF PAGES 67
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

potential outliers
missing points
least-squares
data smoothing
polynomials

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report presents the results of analysis of potential outliers and missing data points in 3-D data. Treatments of isolated and multiple questionable observations (potential outliers and/or missing data points) are suggested for inclusion in the algorithm for smoothing 3-D data using a 7-point least-squares method for fitting polynomials of order three or less.

DD FORM 1473 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ANALYSIS OF POTENTIAL OUTLIER AND MISSING DATA POINTS IN NUWES DATA

by

J.B. Tysver
Associate Professor
Operations Research Department
Naval Postgraduate School
Monterey, California

September 30, 1983

ABSTRACT

This report presents the results of analysis of potential outliers and missing data points in 3-D data. Treatments of isolated and multiple questionable observations (potential outliers and/or missing data points) are suggested for inclusion in the algorithm for smoothing 3-D data using a 7-point least-squares method for fitting polynomials of order three or less.

Keywords: potential outliers, missing points, least-squares, data smoothing, polynomials.

I. INTRODUCTION

The purpose of this report is to present the results of a study of methods of treatment of potential outliers (wild data values) and missing points for inclusion in an algorithm for smoothing of data at NUWES. Potential outliers and missing data points can contaminate both data smoothing (Ref. 1) and geometric analysis of vehicular paths (Ref. 2).

Data used in this investigation were obtained for a single trial run at NUWES. (This run was labeled Trial #2 by this investigator.) Two vehicles (A and B) were involved in this trial. Plots of the horizontal and vertical paths of the two vehicles are shown in Figures 1a,b. Missing points are circled in Figure 1b and denoted by M in data lists. Potential outliers are boxed in figures and denoted by W in data lists.

Data at every eighth scheduled data collection time is missing. In addition, there are other missing data times. Temporary values for these were established as the average of the adjacent values (Ref. 1). Potential Outliers are identified by the use of sequential differences (Ref. 3) with any fourth order difference ($\Delta 4$) having a magnitude of 50 or greater being considered a potential outlier. (The selection of the threshold of 50 is somewhat arbitrary as discussed in Reference 3.)

Data smoothing in this study, and proposed for inclusion in data processing at NUWES, uses the 7-point Least-Squares Polynomial Regression designed for 7 consecutive observations with no missing data (Ref. 1).

A general discussion of the magnitude of the potential outlier and missing point problem is presented in Section II. Their treatment is discussed in Section III.

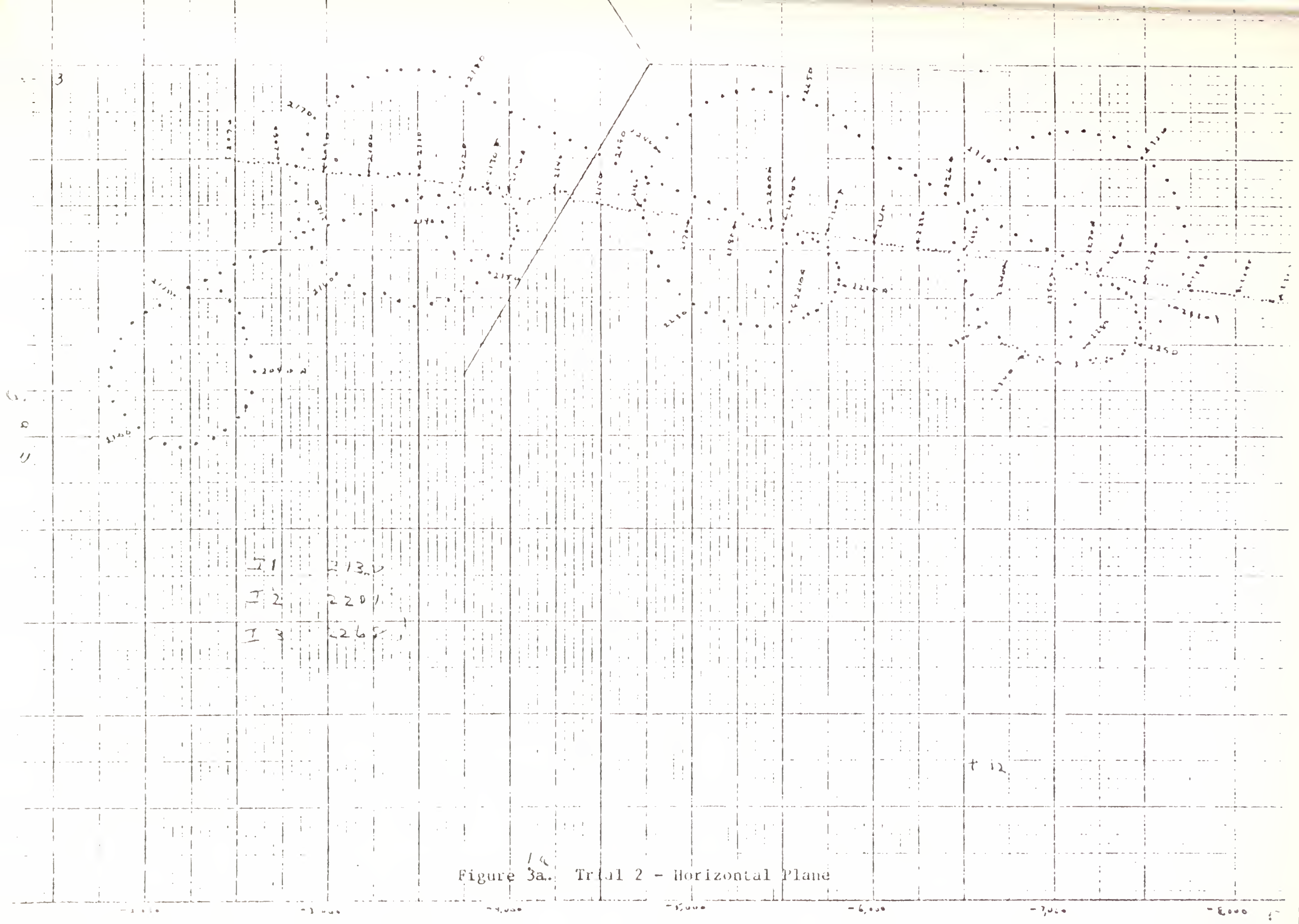


Figure 3a. Trial 2 - Horizontal Plane

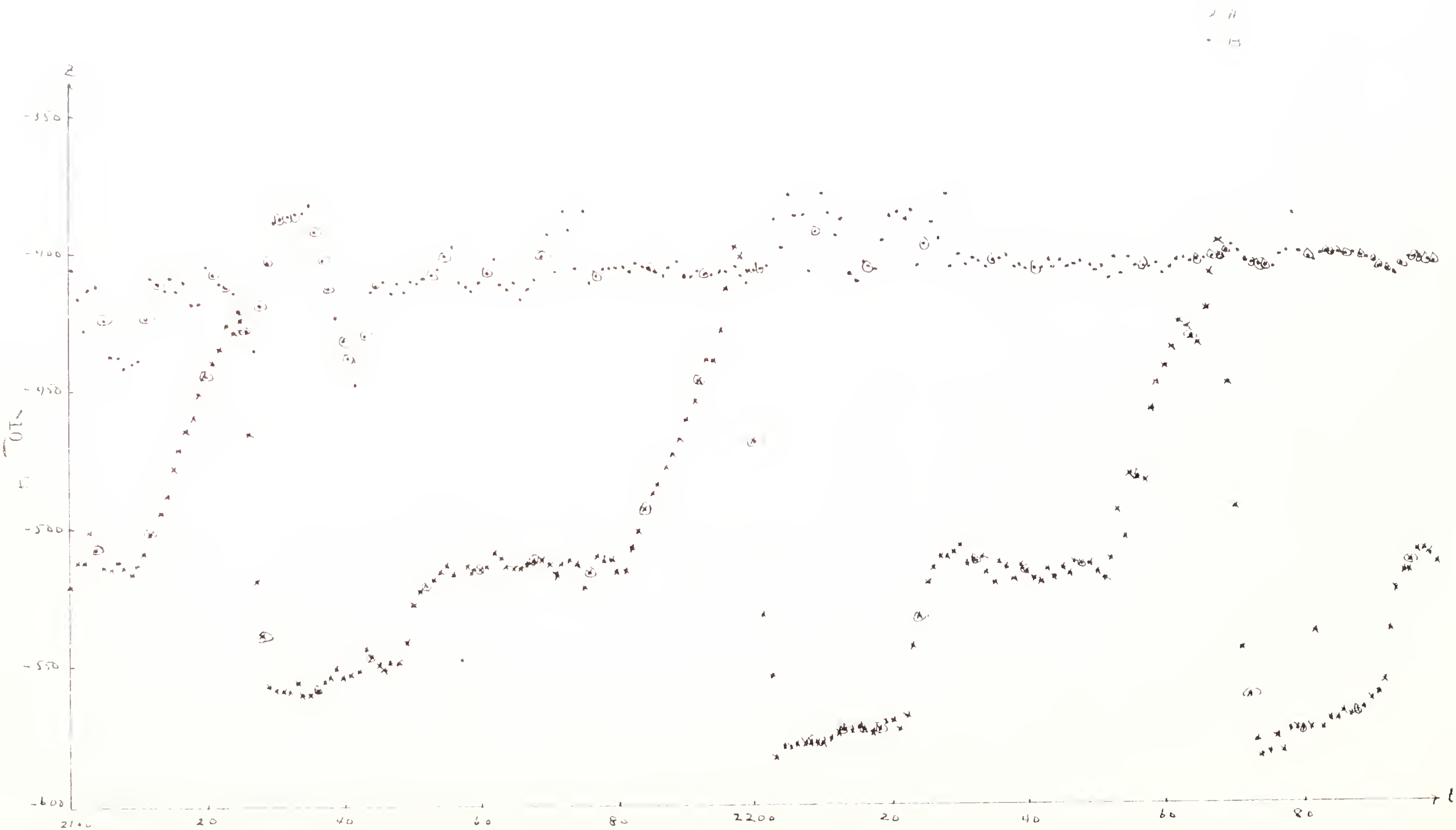


Figure 3b. Trial 2 - Vertical Components

II. GENERAL

The magnitude of the problem of potential outliers and missing points can be demonstrated by the frequency of their occurrence in Trial 2. Observational times for vehicle A were from t = 2084 to t = 2379 and included 296 observational times. These included 36 scheduled missing times (M) and 7 additional unscheduled ones for a total of 43 missing data times. There were also 43 potential outliers (W) in this path with 8 of them designated as both W and M. A summary of the occurrences of wild and missing data is shown below.

Table 1a Vehicle A

Component	W	W & M
x	11	5
y	13	2
z	19	1
x & y	1	0
x & z	1	0
y & z	2	0
x, y, & z	2	0

A similar examination of the path of the vehicle B was also made. Observational times were from t = 2069 to t = 2353 giving 285 observational times. These included 35 scheduled missing times and 48 unscheduled ones for a total of 85. There were 22 potential outliers in this data none of which were also missing data values. This is summarized in Table 1b.

Table 1a Vehicle B

Component	W	W & M
x	5	0
y	5	0
z	12	0
x & y	0	0
x & z	2	0
y & z	2	0
x, y, & z	0	0

Only a brief examination of the extent of wild and missing data was made. Their causes are certainly of concern to data collection personnel but procedures for treatment are of concern for data processing. Some general comments are presented:

(1) There are about seven times as many unscheduled missing data times for the path of vehicle B as there are for that of vehicle A. A cursory examination suggests that these are more prevalent in the path of vehicle B immediately following its approach by vehicle A. Following two of the three approaches in this trial, vehicle A was closer to the nearest tracking array than vehicle B. (Between them? This may be of interest to data collection. These segments of the vehicular paths may be of lesser concern for data processing because they may be of lesser interest to the personnel who are the users of the smoothed data.)

(2) There are about twice as many potential outliers in the path of vehicle A as in that of vehicle B. That their frequency is greater is not unexpected since the vehicle B was doing less maneuvering (ostensibly, on a straight line path). That 8 of the missing values in the path of vehicle A are also designated as potential outliers should not be unexpected. The temporary value inserted for missing values using linear interpolation between adjacent values can be expected to be inconsistent when the actual path is not linear. Note that none of the missing values in the path of vehicle A were also designated as potential outliers.

(3) It is interesting to note the low rate of occurrence of potential outliers in more than one coordinate at the same observation times. For the path of vehicle B these occurred only 6 times and in only two were all three coordinate values indicated as potential outliers. 29 of the 43 potential outliers occurred in one coordinate only. One might be tempted to expect greater multiplicy since any discrepancy in data from the instrumentation arrays is transformed to position coordinates and hence would be expected to contaminate the values of all coordinates at that observational time.

III. TREATMENT OF POTENTIAL OUTLIERS AND MISSING POINTS

A. General

The procedure used in this study (and proposed for data smoothing at NUWES incorporates a 7-point Least-Squares(L-S) polynomial computational routine to treat missing points and potential outliers and, subsequently, for smoothing the rest of the data. Since missing points and potential outliers can contaminate the smoothing of other data points, they should be treated first.

The combination of seven consecutive points for the smoothing routine and the regular scheduling of missing points (every eighth point) complicates the treatment. Operation of chance would dictate that only one time out of eight would a potential outlier or another random missing point be centered in the seven point segment between successive scheduled missing points. A missing point or a potential outlier centered in a seven point segment with no other missing points or potential outliers will be called isolated. These are the easiest to treat. The presence of two or more missing points and/or potential outliers in the same seven point data segment calls for more careful treatment. As discussed in Reference 1, the presence of three such points in a segment should be flagged to indicate to potential users of the smoothed data that the data is of reduced quality.

As discussed in Reference 1, isolated missing points or potential outliers are treated by iterating the 7-point

L-S program replacing the suspect value by the smoothed value at each step and repeating until the smoothed value has a residual error well within the noise of the remaining values in the segment. Since the 'noise' in the NOWES system has a standard deviation of 2 or less for good quality data, the value of 1 has been selected as the magnitude of the residual error for stopping the iteration.

The treatment of multiple missing points and/or potential outliers involves the same procedure with the suspect values replaced by the smoothed values at each step and the smoothing continued until all of the suspect values have residual errors within the specified level (1).

A few missing points and potential outliers in Trial 2 are used to illustrate the smoothing procedure. These are presented in the next section. A 7-point L-S Polynomial program for the TI59 hand-held calculator (see Ref. 1) was used in this treatment.

B. Treatment of Isolated Values

1. An Isolated Missing Point

The isolated missing value selected for illustration of the treatment occurred at times t=2118 in the x coordinate of vehicle A . Data in the vicinity of this point are presented in Figure 2 and Table 2. Also presented in Table 2a are the sequential differences.

Three iterations of the 7-point L-S polynomial smoothing were performed (see Table 2b, columns 2, 3, and 4). The first iteration showed a residual error of $r_0 = -3.76$. Replacing the temporary value $X_0 = 33810.9$ by the smoothed value $X_{01} = 33{,}814.7$ and performing the second iteration showed the residual error reduced to $r_b = -1.23$. Again, replacing by $X_{02} = 33{,}815.9$ (the smoothed value in the second stage) and iterating resulted in the smoothed value $X_{03} = 33{,}816.3$ with the residual error reduced to $r_0 = 0.42$. Since this residual error is less than one in magnitude, the iteration was stopped. Note that the smoothed value X_{02} has a residual error within the specified limits. The third iteration was necessary to establish this. The residual error r_{03} will be even smaller. Since the third estimate X_{03} had to be determining the value of r_{02}, it is accepted as the smoothed value for x at time t = 2128.

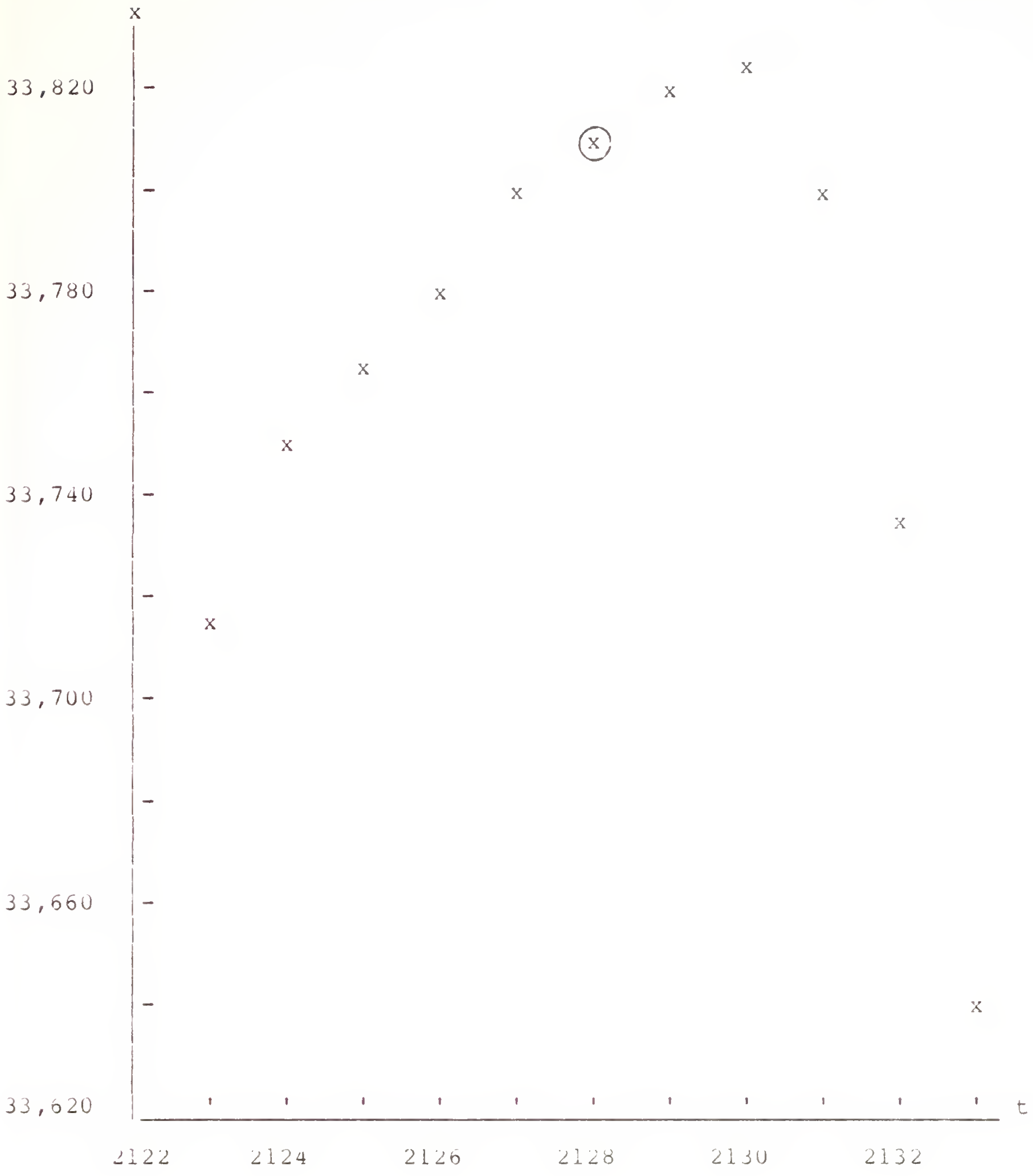


Figure 2 Missing Point (2Ax) t = 2128

Table 2a Isolated Missing Point (2 Ax)

t	x	Before Treatment				After Treatment $x^*=33,816.3$			
		Δ1	Δ2	Δ3	Δ4	Δ1	Δ2	Δ3	Δ4
2123	33,718.5	–	3.8	–	-20.5	–			
–	–	27.3	–	-9.8	–				
2124	33,745.8	–	-6.0	–	8.1				
–	–	21.3	–	-1.7	–				
2125	33,767.1	–	-7.7	–	13.1			–	
–	–	13.6	–	11.4	–		–	11.4	–
2126	33,780.7	–	3.7	–	-19.6	–	3.7	–	-14.1
–	–	17.3	–	-8.2	–	17.3	–	-2.7	–
2127	33,798.0	–	-4.5	–	12.6	–	1.0	–	-9.2
–	–	12.8	–	4.5	–	18.3	–	-11.9	–
2128M	33,810.9	–	0.0	–	-13.7	–	-10.9	–	18.0
–	–	12.9	–	-9.3	–	7.4	–	7.1	–
2129	33,823.7	–	-9.3	–	-18.2	–	-3.8	–	-40.0
–	–	3.6	–	-27.4	–	3.6	–	-32.9	–
2130	33,827.3	–	-36.7	–	29.1	–	-36.7	–	34.6
–	–	-33.1	–	1.7	–		–	1.7	–
2131	33,794.2	–	-35.0	–	13.0				
–	–	-68.1	–	14.7	–				
2132	33,726.1	–	-20.3	–	12.8				
–	–	-88.4	–	27.5	–				
2133	33,637.7	–	7.2	–	-23.4				

Examination of Figure 2 and the first order successive differences ($\Delta 1$) in Table 2α indicates that vehicle A was undergoing a change in path in the vicinity of t = 2128. Sequential differences were recalculated to determine if this change might be indicated by a potential outlier possible at t = 2129 or t = 2130. These values are also presented in Table 2a. The fourth order sequential difference at t = 2129 was increased in magnitude from 18.2 to 40.0 but does not exceed the threshold of 50 so the change in path was not detected by sequential differences.

Because of the change in path (maneuver) of vehicle A, the effect of shifting the segment on the smoothed value was explored. Segments with centers at t = 2126, 2127, 2129, and 2130 were fitted. The smoothed values obtained are presented in Table 2c together with the residual errors r_0, at t = 2128 and the standard deviations (SDR) of the residual errors for the segments. The computations are presented in Tables 2b.

Table 2c - Varying Segments for Smoothing

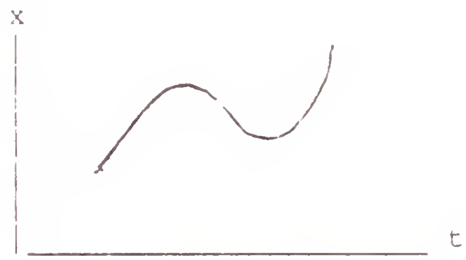
Segment Center	Smoothed x at t = 2128	Residual Error (r_0)	Std. Dev. (SDR)
2126	33810.3	0.57	1.57
2127	33811.7	-0.82	2.06
2128 (M)	33816.3	-0.43	3.55
2129	33813.8	-0.88	3.26
2130	33817.5	-0.81	5.71

Table 2b. Smoothing Missing Point at t = 2128 (2AX)

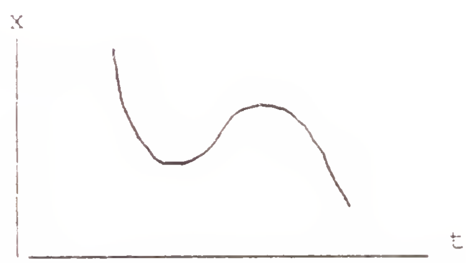
t_i	Segment Center									
	2128	2128(2)	2128(3)	2126	2127	2129(1)	2129(2)	2130(1)	2130(2)	2130(3)
2123	33,718.5									
2124	33,745.8									
2125	33,767.1									
2126	33,780.7			-------						
2127	33,798.0				-------					
2128 M	33,810.9	33,814.7	33,815.9				33,812.9			
2129	33,823.7					-------	-------			
2130	33,827.3							-------	-------	-------
2131	33,794.2									
2132	33,726.1									
2133	33,637.7									
SDR1	17.242	17.774	17.967	5.622	5.285	35.272	35.406	49.019	48.852	48.792
SDR2	9.990	9.750	9.724	2.317	2.090	14.055	13.604	10.934	9.955	9.579
SDR3	4.423	3.638	3.546	1.568	2.065	3.536	3.257	6.469	5.842	5.706
k	3	3	3	3	3	3	3	3	3	3
b_3	-1.2556	-1.2556	-1.2556	0.2556	-0.1472	-1.8667	-1.2778	-1.2778	-1.1667	-1.1167
b_2	-3.5976	-3.7786	-3.8357	-1.2750	-1.2060	-8.0405	-8.1119	-11.7190	-11.7190	-11.7190
b_1	15.9389	15.9389	15.9389	15.2361	14.8841	7.5310	7.0706	-15.3413	-16.4048	-16.8833
b_0	14.3905	15.1143	15.3429	5.1000	4.8238	32.1619	32.4476	46.8762	46.8762	48.8762
x'_{-3}	33,768.4	33,768.0	33,767.9	33,718.8	33,746.7	33,782.0	33,781.8	33,795.9	33,796.7	33,797.0
x'_{-2}	33,778.4	33,779.0	33,779.2	33,745.3	33,764.8	33,794.3	33,795.0	33,814.9	33,816.7	33,817.5
x'_{-1}	33,796.4	33,797.5	33,797.8	33,766.1	33,782.2	33,812.9	33,813.8	33,825.8	33,827.3	33.828.0
x'_0	33,814.7	33,815.9	33,816.3	33,782.9	33,798.2	33,826.6	33,827.1	33,820.9	33,821.4	33,821.7
x'_1	33,825.7	33,826.9	33,827.2	33,797.1	33,811.7	33,824.2	33,824.3	33,792.5	33,792.1	33,792.0
x'_2	33,822.1	33,822.6	33,822.8	33,810.3	33,821.9	33,794.5	33,794.4	33,733.1	33,732.4	33,732.1
x'_3	33,796.2	33,795.8	33,795.7	33,824.0	33,828.0	33,726.4	33,726.5	33,634.9	33,635.2	33,635.4
r_{-3}	-1.27	-0.90	-0.79	-0.33	-0.85	-1.32	-1.13	3.09	1.32	0.98
r_{-2}	2.26	1.72	1.55	0.50	2.33	3.71	2.95	-3.99	-1.80	-0.81
r_{-1}	1.62	0.53	0.19	0.95	-1.54	-1.97	-0.88	-2.06	-3.59	-4.27
r_0	-3.76	-1.23	-0.43	-2.21	-0.19	-2.88	-3.45	6.44	5.87	5.61
r_1	-2.05	-3.13	-3.48	0.87	-0.81	3.10	3.00	1.68	2.06	2.23
r_2	5.20	4.65	4.48	0.57	1.75	-0.34	-0.15	-6.98	-6.31	-6.01
r_3	-2.00	-1.64	-1.52	-0.35	-0.70	-0.30	-0.35	2.83	2.45	2.28

There are several features in these tables that are worthy of comment as follows:

(a) All of the smoothing applications involved a third order polynomial since the standard deviations SDR of the residual errors was smaller for the cubic (SDR3) than for the linear (SDR1) or the quadratic (SDR2). The cubic polynomical used to fit the data segment with center at t = 2126 is of the form



since the coefficient (b_3) of the third order term is positive ($b_3 > 0$). For all other segment centers the cubic is of the form



since the coefficient b_3 is negative. These results suggest that the data segments centered at t = 2127 to t = 2130 included positions in the maneuver.

(b) The smoothed values for x at time t = 2128 vary more than 7 units depending upon the data segment used for the smoothings. The question now arises of which smoothed value provides the best estimate of the x coordinate of vehicle A at time t = 2128. The residual error at t = 2128 provide no help here since it could be reduced to zero by repeated iteration.

Note that, as discussed in the smoothing of the segment centered at t = 2128, the residual error is the difference between the temporary value before the last iteration and the smoothed value after that iteration and hence does not represent an error in the smoothed value. It should be noted also that further iteration to reduce the residual error at t = 2128 will only produce small reductions in the standard deviations of the residual errors of the segments since the purpose of the iterations is to reduce the residual error at that point to a value well within the residual errors at the other points in the same data segments (i.e., small contribution with respect to the 'noise' in the segments).

(c) Of greater use for selecting the most appropriate data segment, and consequently, of the most appropriate estimate for x at t = 2128, are the values of the standard deviation of the residual errors (SDR, in Table 2c). The standard deviation of the residual errors is used in establishing confidence intervals for the actual value of the dependent variable (x) with smaller values producing narrower confidence intervals (Ref.1). The data segment centered at t = 2126 had the smallest standard deviation and hence could be considered to give the preferred estimate.

The variation of the width of the confidence interval with the degree of polynomial used to fit the data segment and with the location of the missing data point within the segment has not been fully explored. The first degree polynomial was treated in Reference 1 but similar expressions for confidence intervals when second and third order polynomials are used needs further development.

(d) There should be some concern about the effect of the change in vehicular path on the smoothing of the data. This change occurred in the vicinity of times t = 2128 or 2129. (See Fig.2)

Note, one possible explanation for the increase in the value of SDR as the center of the data segment is shifted is an increase in the 'noise' level in the observations. Another is the inability of the third order polynomial to represent the actual vehicular path adequately. In order to avoid the latter possibility it would appear desirable to avoid smoothing data with segments including rapid changes in vehicular paths. Referring, again, to Figure 2, it can be seen that a rather abrupt change in the vehicular path is apparent at time t = 2130 but that the observation at time t = 2129 appears to be consistent with the preceding values. Thus exclusion of the observation at t = 2130 from the segment would lead to using the 7-point data segment centered at t = 2126. Further, this same segment should also be used for subsequent smoothing of data values at times 2127 and 2129 instead of using data segments centered at those times. (Note that this suggestion of using the data segment centered at t = 2126 to smooth the value for the missing point at t = 2128 is in accord with the discussion in comment c above.)

(e) The major guidelines in developing a data smoothing algorithm for use at NOWES included:

(1) the resulting data smoothing program should be as fully automated as possible, and

(2) the resulting data smoothing program should be as simple and short as possible.

These two guidelines are contradictory when it comes to treatment of changes in vehicular paths.

It could be very awkward to construct subroutines to implement automatic identification of the times of changes in vehicular paths. On the other hand, manual screening of the data to identify such times would reduce the level of automation.

Fortunately there is another source of information that could be made available to provide this information. This is the internal control information collected from the vehicles. It is strongly recommended that this source of information be explored with the intent of including it with the data to be smoothed.

2. An Isolated Potential Outlier

As discussed in Section IIIA, isolated potential outliers are rare. One occurrence in the trial used in this report was the y coordinate of vehicle A at time $t = 2268$. The data in the vicinity of the potential outlier is presented in Table 3a together with the sequential difference ($\Delta 4$). At $t = 2268$ $\Delta 4 = 75.1$ which exceeds the selected threshold magnitude of 50 and hence the y value at $t = 2268$ is indicated as a potential outlier. A plot of the data is also presented in Figure 3.

Treatment of a potential outlier is the same as that for an isolated missing point. Four iterations were required to ensure that the smoothed value to be used as a replacement for the potential outlier was consistent with the other six values in the data

Figure 3 Isolated Potential Outlier 2AY t = 2268

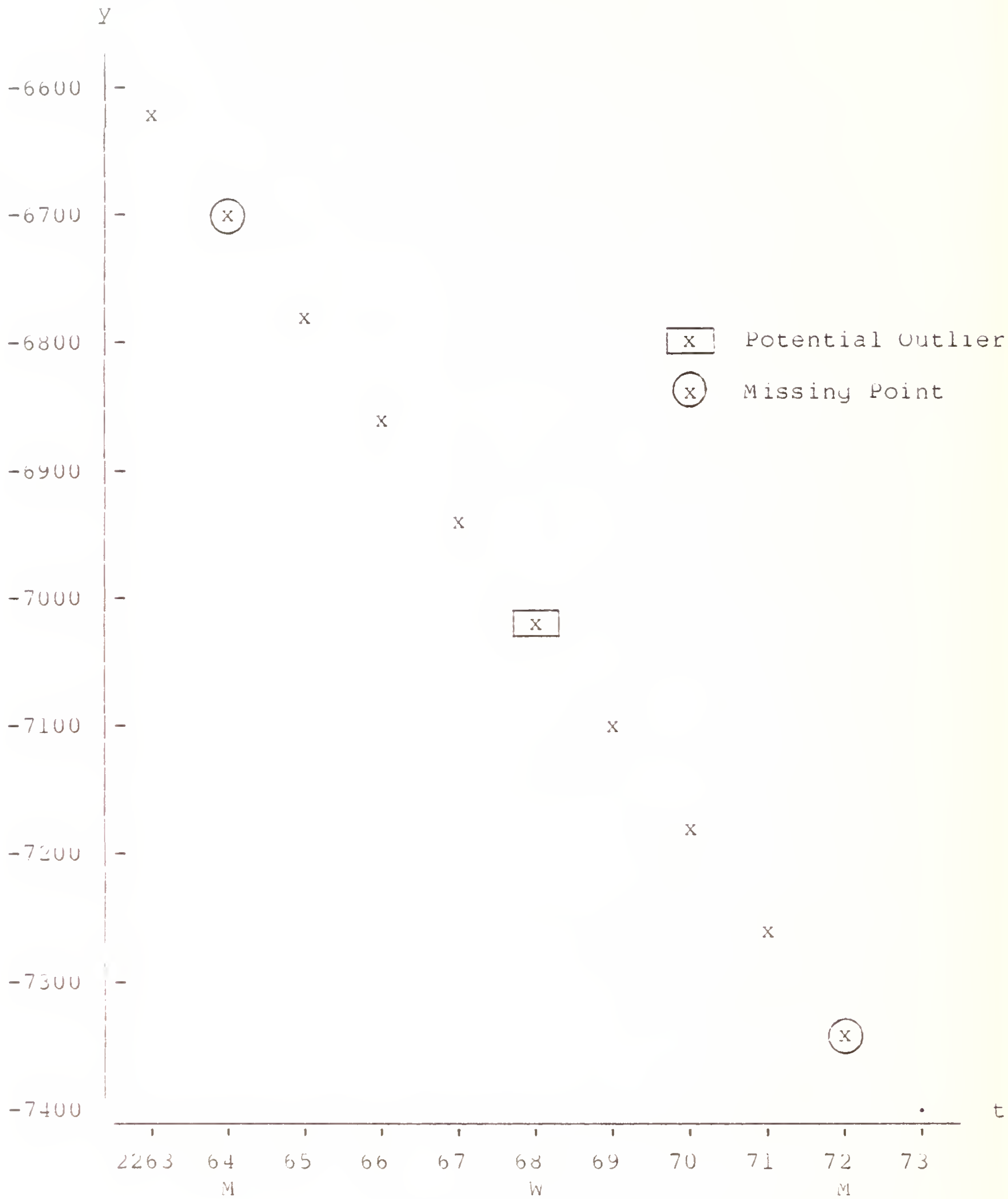


Table 3a

SEQUENTIAL DIFFERENCES FOR AN ISOLATED POTENTIAL OUTLIER

t	y	Before Treatment $\Delta 4$	After Treatment $\Delta 4$
2263	-6627.6		
--	--		
2264M	-6704.2		
--	--	--	
2265	-06781.8	18.1	
--	--	--	--
2266	-6868.8	-9.1	6.3
--	--	--	--
2267	-6957.5	36.1	-25.5
--	--	--	--
2268	-7048.8	<u>-75.1W</u>	1.9
--	--	--	--
2269	-7107.5	42.6	11.8
--	--	--	--
2270	-7173.5	-3.2	12.2
--	--	--	--
2271	-7244.1	5.6	
--	--	--	
2272M	-7319.8		
--	--		
2273	-7395.5		

Treatment: Observed value y = 7048.8 at t = 2268 replaced by smoothed value y = 7033.4

segment, i.e., that the residual error of the smoothed value was less than the specified magnitude of one. The fourth iteration was required to determine whether the smoothed value obtained in the third iteration satisfied this criterion. As in the treatment of an isolated missing point (Section III B1), the smoothed value $x_4 = 7033.4$ established in the fourth iteration was selected as the replacement for the observed value $\bar{x}_1 = 7048.8$ and will have a residual error about the fitted curve which is less than $r_3 = 0.39$. The iterations conducted on a TI59 are presented in Table 3b.

There are several features of this treatment which are worthy of comment.

(a) Sequential differences were recalculated after replacing the potential outlier. These are presented in the right hand part of Table 3a. The fourth order difference at t = 2268 has been reduced in magnitude from 75.1 to 1.9 and elimination of the contamination of the fourth order differences at the adjacent times has also reduced their magnitude.

(b) It is of some interest to note that in the first two iterations a second order polynomial (parabola) provided the best fit (smallest SDR) but a third order polynomial (cubic) gave a slightly better fit in the last two iterations.

Table 3b Isolated Potential Outlier (2AY)

	Smoothing Iteration			
	y_i	$y_{i(2)}$	$y_{i(3)}$	$y_{i(4)}$
2265	-6781.8			
2266	-6866.8			
2267	-6957.5			
2268 W	-7048.8	-7038.4	-7034.9	-7033.8
2269	-7107.5			
2270	-7173.5			
2371	-7244.1			
SDR 1	14.157	11.141	10.333	10.109
SDR 2	6.947	3.504	2.877	2.804
SDR 3	7.819	3.627	2.797	2.697
R	2	2	3	3
b_3	--	--	-0.2111	-0.1111
b_2	3.1036	2.6083	2.4417	2.3893
b_1	-76.6536	-76.6536	-75.1758	-75.1758
b_0	-12.4143	-10.4333	9.7667	-9.5571
y'_{-3}	-6780.5	-6781.5	-6780.6	-6780.7
y'_{-2}	-6872.7	-6871.2	-6872.0	-6871.8
y'_{-1}	-6958.7	-6955.7	-6956.0	-6955.6
y'_0	-7038.4	-7034.9	-7033.8	-7033.4
y'_1	-7112.0	-7108.0	-7106.7	-7106.4
y'_2	-7179.3	-7177.8	-7176.1	-7175.9
y'_3	-7240.4	-7241.4	-7248.0	-7243.1
r_{-3}	-1.28	-0.29	-1.22	-1.12
r_{-2}	3.89	2.41	3.17	3.02
r_{-1}	1.16	-1.81	-1.55	-1.86
r_0	-10.39	-3.45	-1.12	0.39
r_1	4.46	1.49	-0.77	-1.08
r_2	5.81	4.32	2.55	2.40
r_3	-3.66	-2.67	-1.07	-0.96

c) The reduction in the residual error in the potential outlier and in the standard deviation (SDR) of the residual errors of the data segment are worthy of notice. The residual error was $r_0=-10.39$ for the potential outlier and the standard deviation of the residual errors (the differences between observational values and smoothed values) was SDR2 = 6.95. The third iteration of smoothing replaced the potential outlier value of $X_0= -7048.8$ by $x_3 = -7033.8$ which as a residual error of $r_3=-0.4$ and the standard deviation of the residual errors was SDR3 = 2.70 (established in the fourth iteration).

d) The magnitudes of all of the residual errors when a data segment is smoothed is of some concern. This is represented by the value of the SDR which was somewhat larger in all but one of the data segments examined in the previous subsection (IIIB1) where an isolated missing point was considered. There is always some reservations in the mind of this investigator (and should be in the mind of any potential user of the smoothed data) whether a larger value of the SDR is caused by inadequacy of the model (polynomials of order three or lower) or an increase in the level of noise in the data.

Inadequacy of the model is not limited to major changes in a vehicular path as apparently occurred in the missing point example but could be produced by fish-tailing (snake action) for vehicular control or minor corrections in attack path. A higher data rate would improve the smoothing capabilities for following such higher frequency path variations by allowing use of longer path segments and/or higher order polynomials as well as improved smoothing capabilities even when such path anomalies were not present.

The presence of an unscheduled missing point or of an outlier when the SDR for the residual error is large should not be unexpected. It should serve as an indication that the position location system is having difficulty in obtaining consistent data on the vehicular path.

The inability of the smoothing procedure to distinguish between inadequacy of model and noise as the cause for larger values of the SDR should be recognized as a different kind of inaccuracy of the model. In the development of the Least-Squares Model it was assumed that the noise components of the observed values were independent. Any persistence in the noise component is thus treated as a portion of the actual path component. As an extreme example, any constant portion of the noise component that persists over an entire data segment will result in a bias in the smoothed path, i.e., in an offset of the smoothed path from the actual path of the vehicle.

3. An Isolated Missing Point/Potential Outlier

The fact that the temporary replacement of an isolated missing point by the average of the values at the adjacent points can produce a value which is identified by the sequential differences as a potential outlier is illustrated by the x-coordinate of vehicle A at time t = 2136. The data segment and the sequential differences are presented in Table 4a and sketched in Figure 4. The four smoothing iterations are shown in Table 4b.

The treatment here is not different from that of an isolated missing point or a potential outlier. It is included in this report to illustrate that the temporary replcement of a missing point by the average of the adjacent points is actually using a 2-point straight line fit and hence may be substantially different from the actual value of the component when the vehicle is not traveling in a straight line.

One other side comment that may be of interest is the magnitudes of the SDRs in the second and third smoothing iterations in comparison with the actual values of the residual errors. The SDR of the residual errors is larger than any of the residual errors in these iterations. This is a consequence of using

$$SDR = [\frac{1}{n-4} \sum (r_i - \bar{r})^2]^{\frac{1}{2}} = [\frac{1}{n-4} \sum r_i^2]^{\frac{1}{2}}$$

instead of the root-mean-square

Figure 4 ISOLATED MISSING POINT/POTENTIAL OUTLIER

2Ax t = 2136

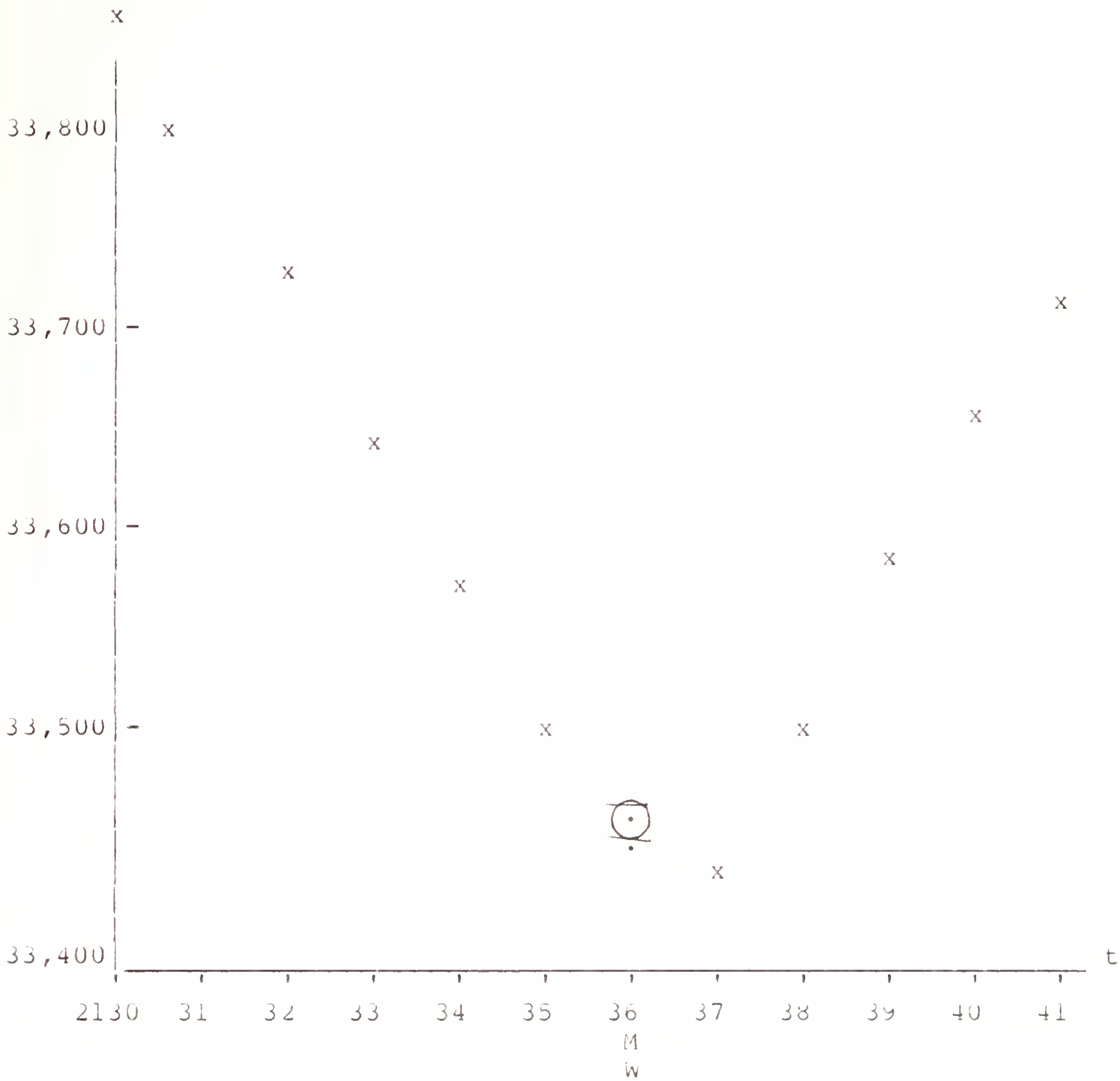


Table 4a

ISOLATED MISSING POINT/POTENTIAL OUTLIER 2AX

t	x	Before Treatment Δ4	After Treatment x*=33,452.2 Δ4
2131	33,794.2		
--	--	--	
2132	33,726.1	12.8	
--	--	--	
2133	33,637.7	-23.4	
--	--	--	--
2134	33,556.5	34.3	20.1
--	--	--	--
2135	33,486.6	-88.2	-31.6
--	--	--	--
2136M	33,466.5	108.7	23.2
--	--	--	--
2137	33,446.3	-82.5	-25.5
--	--	--	--
2138	33,485.1	4.8	-9.4
--	--	--	--
2139	33,559.3	-0.6	
--	--	--	
2140	33,650.2	-5.9	
--	--		
2141	33,738.5		

$$RMS = \left[\frac{1}{n} \sum r_i^2 \right]^{\frac{1}{2}}$$

when a cubic polynomial is fitted to a data segment of n=7 observations. The unbiased estimator SDR of the standard deviation for the noise component is considerably increased over the RMS value because the data segment is so short. An increase in the data rate to increase n is desirable. Note that a kth order polynomial would require a divisor of n-(k+1) since it would involve k+1 coefficients. Thus a substantial increase in n (e.g., doubling the data rate) would permit some increase in the order of polynomials considered for fitting the data segment without making the value of SDR unrepresentative of the residual errors.

Table 4b Isolated Missing Point/Potential Outlier (2Ax)

t_i	Smoothing Iteration			
	x_i	$x'_{i(2)}$	$x'_{i(2)}$	$x'_{i(4)}$
2133	33,637.7			
2134	33,556.5			
2135	33,486.6			
2136 M,W	33,466.5	33,456.8	33,453,6	33,452,5
2137	33,446.4			
2138	33,485.1			
2139	33,559.3			
SDR 1	64.974	66.665	67.266	67.477
SDR 2	9.428	7.599	7.372	7.346
SDR 3	7.545	3.922	3.294	3.216
b	3	3	3	3
b_3	0.925	0.9250	0.925	0.925
b_2	15.7179	16.1798	16.3321	16.3845
b_1	-21.4143	-21.4143	-21,4143	-21.4143
b_0	-62.8714	-64.7190	-65.3286	-65.5381
x'_{-3}	33,637.6	33,638.5	33,638.8	33,638.9
x'_{-2}	33,555.1	33,553.8	33,553.3	33,553.1
x'_{-1}	33,493.1	33,490.3	33,489.4	33,489.1
x'_0	33,456.8	33.453.6	33,452.5	33,452.2
x'_1	33,452.1	33,449.3	33,448.4	33,448.1
x'_2	33,484.3	33,482.9	33,482.4	33,482.3
x'_3	33,559.0	33,560.0	33,560.3	33,560.4
r'_{-3}	0.13	-0.80	-1.10	-1.20
r'_{-2}	1.36	2.74	3.20	3.36
r'_{-1}	-6.45	-3.68	-2.76	-2.45
r'_0	9.66	3.19	1.06	0.32
r'_1	-5.77	-3.00	-2.09	-1.77
r'_2	0.31	2.20	2.66	2.81
r'_3	0.26	-0.66	-0.96	-1.07

C. Treatment of Multiple Values

1. General Considerations

When more than one missing point and/or potential outlier occur in the same 7-point data segment the selection of the appropriate treatment is more difficult. Treatment of data segments containing 3 or more values which are either missing points or potential outliers require additional considerations and will be postponed until the next section (Sect. D). Only occurrences of two such values will be examined here.

Treatment of two such suspect values must take into consideration the differences in the nature of suspect values as well as their location in a 7-point segment. There are three possible procedures:

a) Smooth first one using iterations as necessary, then the other using the smoothed value for the first. It would appear advisable to resmooth the first again after the second is smoothed. The question arising here is which value should be smoothed first. In the case of two potential outliers it would appear reasonable to smooth first on the one with the largest fourth order difference ($\Delta 4$) as representing the greater potential contaminator. In the case of a potential outlier and a missing point it would appear reasonable to smooth the potential outlier first for the same reason. In the case of two missing points this reason is not pertinent and a reasonable procedure would be to smooth the one that occurred first in time for computational simplicity.

(b) Alternate smoothing iterations centered on first one time then the other, continuing the iterations until the residual errors of both are within the prescribed limits. This procedure requires more computational effort since the 7-point segments shift between each iteration. There is also the possibility that, because different data segments are involved, both residual errors cannot be reduced to the prescribed level simultaneously.

(c) Simultaneous smoothing of the two values using a single 7-point segment. The question here is where the segment should be centered. This selection should take into consideration the quality of the resulting smoothed values.

As discussed in Reference 1, the quality of a smoothed value can be expressed in terms of the width of the confidence interval for the actual value at any time. This confidence interval is of the form

$$CI_{1-\alpha}(x_t) = [\hat{x}(t) \pm t_{\alpha/2} s_r \sqrt{\frac{1}{n} + \frac{(t-\bar{t})^2}{\sum(t_i-\bar{t})^2}}]$$

$$= [\hat{x}(t) \pm t_{\alpha/2}\, s_r \sqrt{\frac{1}{7} + \frac{t'^2}{28}}] \qquad (1)$$

when the t values are translated to $t' = -3, -2, -1, 0, 1, 2, 3$ for the 7-point segment when the fitting polynomial is linear. (The comparable forms for quadratic and cubic polynomials has not been explored.) Thus the values to be smoothed should be as close to the center of the segment as possible since the confidence interval will be shortest when $t' = 0$.

Situations in which adjacent points are both potential outliers are unlikely occur with the identification procedure specified (Ref. 3) since only the point having the largest fourth order difference ($\Delta 4$) exceed-ing the prescribed level (50) has been so labeled. (To guard against outliers close to each other, sequential differences should be recalculated whenever a potential outlier has been smoothed.) This is illustrated in Section III C 2.

Situations in which adjacent points consist of a potential outlier and a missing point should be treated simultaneously using the data segment centered on the potential outlier since it contaminates the temporary value assigned for the missing point. This is illustrated in Section III C 3.

For situations with adjacent missing points, simultaneous smoothing is again recommended. It is, however, ambiguous as to which one should be used as the center of the data segment used for smoothing. This is examined in Section III C 4 for one such occurrence in the trial run.

When a missing point occurs in the 7-point data segment centered at a potential outlier but is not adjacent to it, there is some question as to whether it should be smoothed simultaneously or subsequent to the treatment of the potential outlier. This has not been examined but, on the principle of making the associated confidence interval as short as possible, the latter would appear preferable.

For two missing points in the same 7-point data segment which are not adjacent, the treatment can be different depending on their separation. If they are separated by only one point the possibility of simultaneous smoothing using that point as the center of the data segment would be advantageous from a computational viewpoint and would not cause a substantial increase in the width of the confidence interval. This can be seen in the factor

$$\frac{1}{n} + \frac{t^2}{\sum t_i^2} = \frac{1}{7} + \frac{1}{28}$$

for $t = \pm 1$ in Equation (1). This situation is examined in Section III C 5. The situation when two missing points are separated by two other points is also explored in Section III C 5.

III. C 2 Two Potential Outliers

As discussed in Reference 3, a large fourth order sequential difference ($\Delta 4$) indicating a potential outlier is typically accompanied by large $\Delta 4$'s for the adjacent values but with opposite signs. These may also exceed the specified threshold but should not initially be labeled as potential outliers. This is illustrated in the Table 5a and Figure 5. Note that the $\Delta 4$'s at times 2212, 2213, and 2214 all exceed the threshold 50, that their signs alternate, and that the magnitude of $\Delta 4$ at 2213 is largest. Only the value of z at $t = 2213$ should be considered a potential outlier. Smoothing this value (Table 5b) and recalculating the $\Delta 4$'s verifies that the values at $t = 2212$ and 2214 are not potential outliers and that their $\Delta 4$'s were contaminated by the designated outlier at $t = 2213$.

If the second calculation indicates another potential outlier in the vicinity of the first one, then the suggested procedure for smoothing can depend on their separation. The data from Trial Run #2 was not examined to see whether this occurred. Treatment for such a situation is the same as that for two missing points and will be presented in Section III C 3.

Table 5a

Potential Outlier at $t = 2213$ (2BZ)

t_i	t_i	z_i	$\Delta 4$	$\Delta 4^*$ ($z_0 = 402.3$)
-4	2209 M	-393.3		
-3	2210	-379.1	42.3	
-2	2211	-386.5	5.5	-19.0
-1	2212	-394.8	-98.2	-2.2
0	2213 W	-377.8	<u>144.9</u>	-2.1
1	2214	-407.5	-88.8	-9.2
2	2214	-411.0	-2.6	-27.1
3	2216	-404.2	26.7	
4	2217	-405.6		

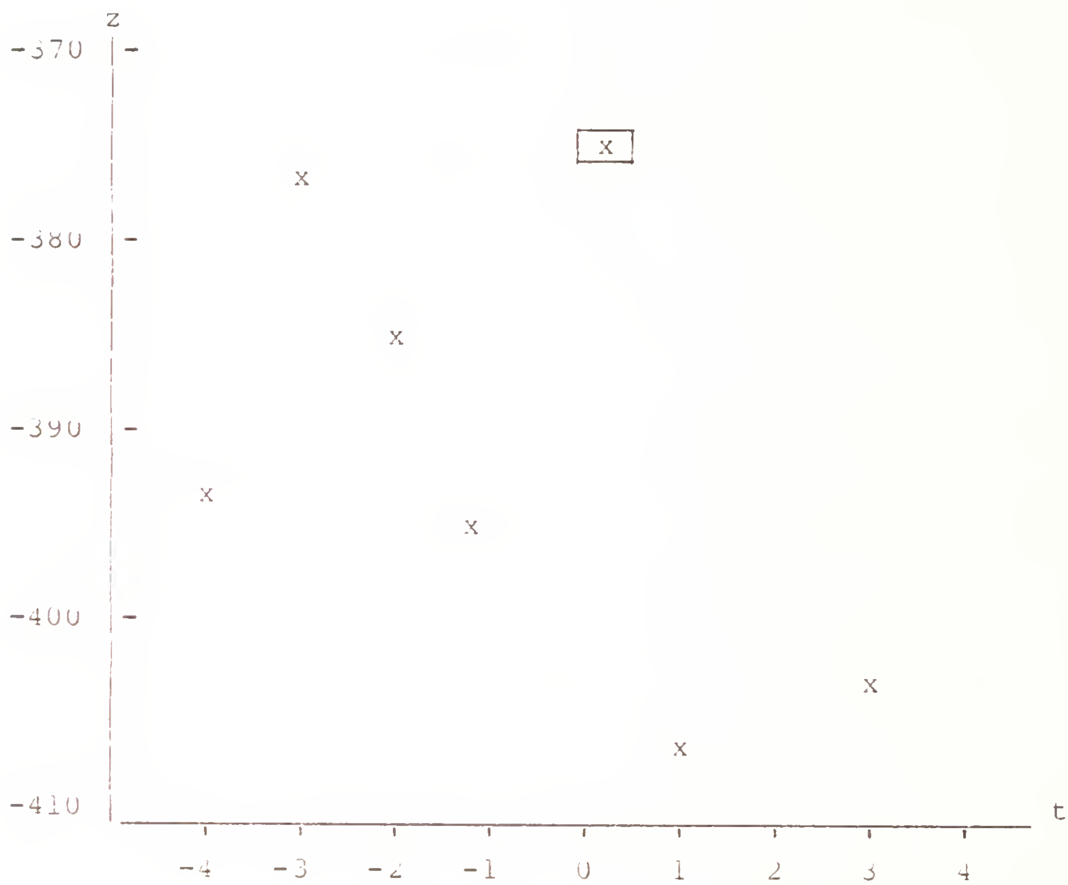


Figure 5 Potential Outlier at $t = 2213$ (2BZ)

Table 5b Smoothing Potential Outlier in 2BZ at t = 2213

t	t^1	z_i	$z_{i(2)}$	$z_{i(3)}$	$z_{i(4)}$
2210	-3	-379.1			
2211	-2	-3865.5			
2212	-1	-394.8			
2213W	0	-377.8	-394.4	-400.0	-401.9
2214	+1	-407.5			
2215	+2	-411.0			
2216	+3	-404.2			
SDR1		9.435	5.093	5.096	5.332
SDR2		10.548	4.320	2.854	2.636
SDR3		11.842	4.093	1.651	1.065
R		1	3	3	3
b_3		--	.33611	.33611	.33611
b_2		--	.80952	1.0762	1.1667
b_1		-4.8929	-7.2456	-7.2456	-7.2456
b_0		-394.41	-3.2381	-4.3048	-4.6667
z'_{-3}		-379.7	-380.1	-379.5	-379.4
z'_{-2}		-384.6	-385.0	-385.8	-386.1
z'_{-1}		-389.5	-392.3	-393.9	-394.4
z'_0		-394.4	-400.0	-401.9	-402.5
z'_{+1}		-399.3	-406.1	-407.7	-408.3
z'_{+2}		-404.2	-408.6	-409.4	-409.7
z'_{+3}		-409.1	-405.4	-404.9	-404.7
r_{-3}		.64	.98	.44	.26
r_{-2}		-1.87	-1.52	-.72	-.45
r_{-1}		-5.28	-2.50	-.90	-.35
r_0		16.61	5.62	1.89	.62
r_1		-8.19	-1.38	.22	.77
r_2		-6.8	-2.41	-1.61	1.34
r_3		4.89	1.20	.67	.49

III. C 3 A Potential Outlier and a Missing Point

When a missing point is adjacent to a potential outlier its temporary value is the average of the potential outlier and the neighboring value on its other side. It would appear reasonable for this situation to smooth the two values simultaneously using the data segment centered on the potential outlier. An example of this occurred at times 2175(W) and 2176(M) in the x coordinate of vehicle A in Trial 2. The appropriate data is presented in Table 6a and Figure 6. The TI 59 calculator output is shown in Table 6b. Sequential differences were recalculated since a potential outlier was present and the $\Delta 4^*$'s are also presented in Table 6a.

If a missing point occurs in the data segment centered at a potential outlier but is not adjacent to it, then simultaneous smoothing may not be appropriate. Note that the factor

$$\frac{1}{7} + \frac{t'^2}{28}$$

in determining the width of the confidence interval is

$\frac{1}{7} + \frac{4}{28} = \frac{2}{7}$ for $t = \pm 2$ and $\frac{1}{7} + \frac{9}{28} = \frac{13}{28}$ for $t = \pm 3$. Thus the width of the confidence interval at these times is substantially increased. It would appear reasonable in such situations to smooth the potential outlier first, then smooth the missing point, and, if desired, to resmooth the potential outlier.

Table 6a Adjacent Potential Outlier and Missing Point

t^1	t	X_i	Δ4	$\Delta 4^*$ (X_0=34,463.4) (X_1=34,477.1)
-3	2172	34,341.7	-30.8	
-2	2173	34,395.5	17.7	-22.0
-1	2174	34,436.6	-36.1	2.8
0	2175 W	34,472.3	63.1	-3.7
+1	2176 M	34,473.8	-45.6	10.1
+2	2177	34,475.2	1.8	-20.5
+3	2178	34,465.3	21.4	24.7
+4	2179	34,434.5	-9.6	

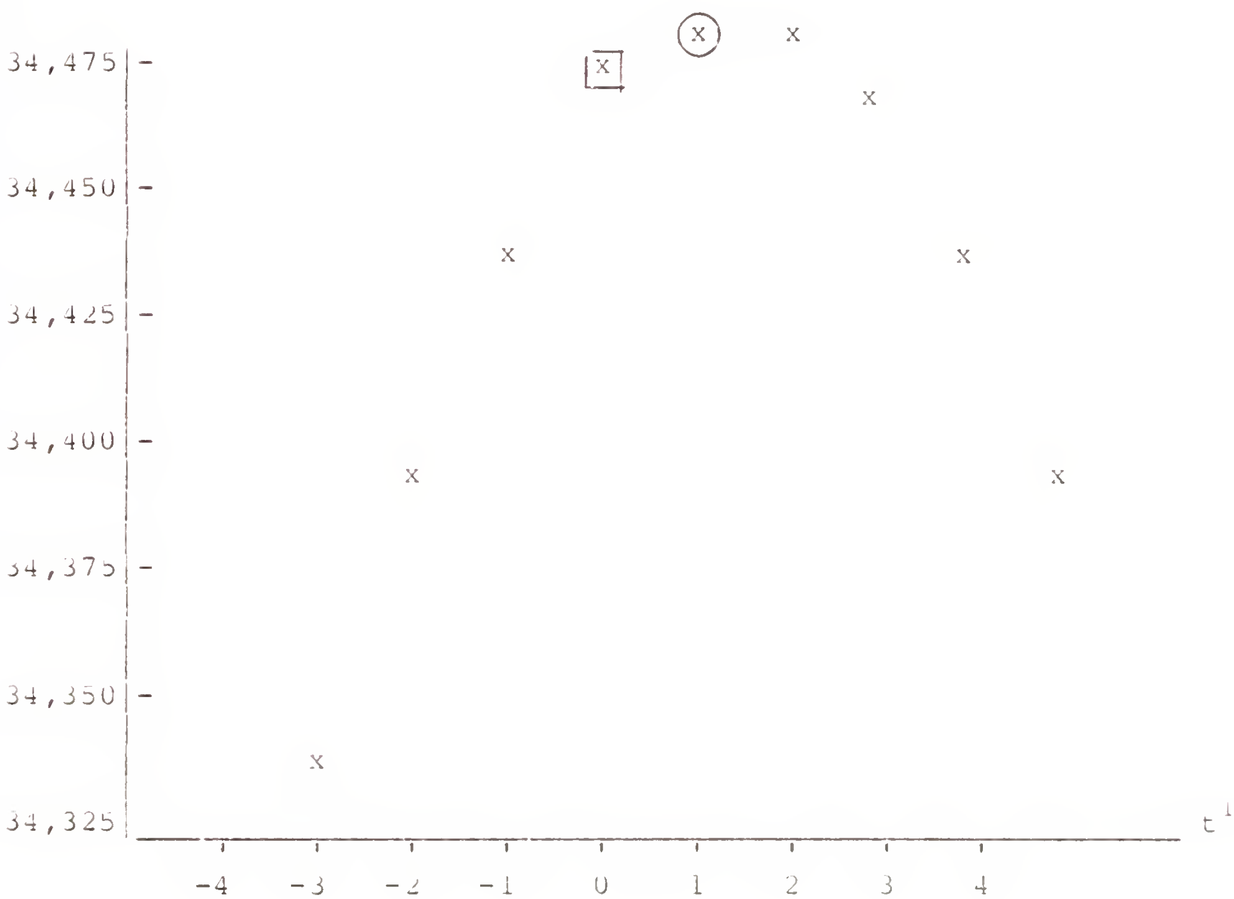


Figure 6 Adjacent Potential Outlier and Missing Point

Table 6b Smoothing Adjacent Potential Outlier and Missing Point

t	t^1	X_i	$X_{i(2)}$	$X_{i(3)}$
2172	-3	34,341.7		
2173	-2	34,395.5		
2174	-1	34,436.6		
2175W	0	34,472.3	34,465.1	34,464.0
2176M	+1	34,473.8	34,478.4	34,477.6
2177	+2	34,475.2		
2178	+3	34,465.3		
2179	+4	34,434.5		
SDR1		28.868	27.867	27.525
SDR2		4.665	1.546	1.289
SDR3		5.150	1.715	1.323
R		2	3	3
b_3		--	--	--
b_2		-6.9690	-6.7905	-6.7095
b_1		20.2643	20.4286	20.4
b_0		27.8762	27.1619	26.8381
X'_{-3}		34,341.6	34,341.6	34,341.8
X'_{-2}		34,396.7	34,395.4	34,395.8
X'_{-1}		34,437.8	34,436.8	34,436.3
X'_0		34.465.1	34,464.0	34,463.4
X'_{+1}		34,478.4	34,477.6	34,477.1
X'_{+2}		34,477.7	34,477.7	34,477.4
X'_{+3}		34,463.1	34,464.2	34,464.2
r_{-3}		0.14	0.11	-0.11
r_{-2}		-1.17	-0.47	-0.26
r_{-1}		-1.24	-0.17	0.31
r_0		7.22	1.11	0.60
r_1		-4.57	0.77	0.51
r_2		-2.53	-2.49	-2.16
r_3		2.15	1.14	1.09

III C 4 Two Missing Points

Some exploration will be presented here of the effects of different treatments of two suspect values when they are adjacent, separated by a single value, or separated by two values. First, consider a situation with two adjacent missing points with no other suspect values in the 7-point data segment centered on either of them. It would appear reasonable to use simultaneous smoothing using the data segment centered on either one. An example of this in Trial 2 data is presented in Table 7a and Figure 7. The outputs of the TI 59 calculator smoothing are shown in Table 7 using the data segment centered at t= 2352. This example is of some interest since the fitted cubic polynomial changes drastically with the shift of one unit in the data segment location. This is indicated by the coefficient b_3 of t^3 which is positive when the segment is centered at t = 2352 and is negative when the segment is centered at t = 2353 (See Section III B1). In spite of this difference in the fitting cubic polynomials, the smoothed values do not differ drastically from each other or from the temporary values initially used. Whether the differences in the smoothed values shown in Table 7 are of concern to potential users of the smoothed data is uncertain. If it is not, then simultaneous smoothing can be used with either missing point at the center of the data segment.

Smoothing of these points using simultaneous smoothing but alternating the center between the missing points at successive iterations has not been explored since only one smoothing step

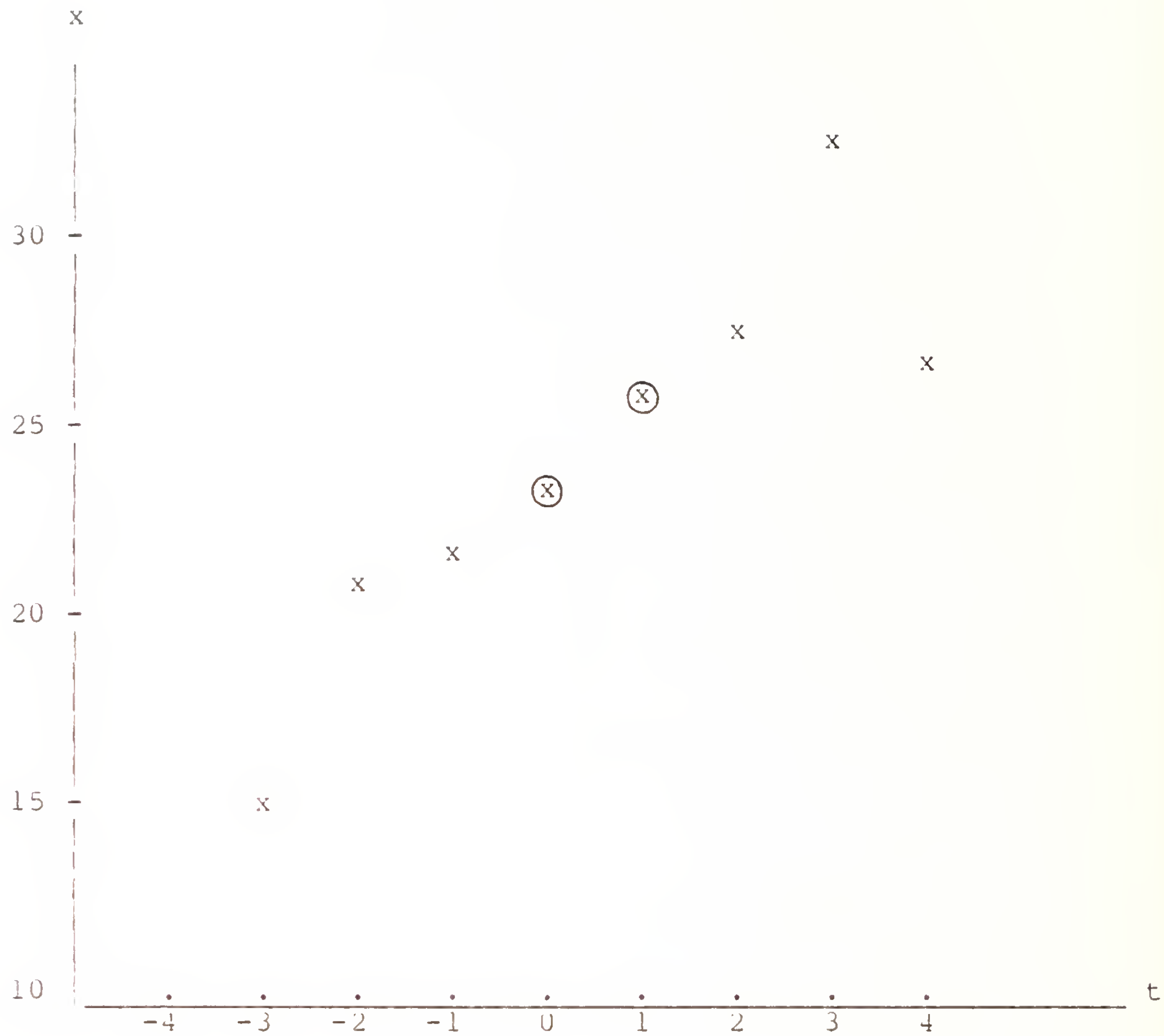


Figure 7. Adjacent Missing Points.

Table 7 Smoothing Adjacent Missing Points

t	t'_i	$X_i^{(1)}$	t'_i	$X_i^{(2)}$
2349	-3	33,015.3		
2350	-2	33,021.1	-3	
2351	-1	33,022.0	-2	
2352M	0	33,024.0	-1	
2353M	+1	33,026.1	0	
2354	+2	33,028.1	1	
2355	+3	33,033.0	2	
2356		33,027.0	3	
SDR1		1.334		2.443
SDR2		1.490		2.499
SDR3		.736		1.904
R		3		3
b_3		0.1833		-0.2556
b_2		0.0143		-0.1405
b_1		1.2595		3.3432
b_0		-0.0571		0.9819
X'_{-3}		33,015.6		33,021.6
X'_{-2}		33,020.2		33,021.2
X'_{-1}		33,022.7		33,023.5
X'_0		33,024.2		33,026.9
X'_{+1}		33,025.6		33,029.7
X'_{+2}		33,028.2		33,030.6
X'_{+3}		33,033.0		33,027.9
r_{-3}		-0.27		
r_{-2}		0.86		
r_{-1}		-0.74		
r_0		-0.17		
r_1		0.47		
r_2		-0.11		
r_3		-0.03		

(1) Segment center at t = 2352

(2) Segment center at t = 2353

using either center brings residual errors for both replacement values within the desired level ($|r_i| < 1$). Such alternation of data segment centers could require substantial computational effort using the TI59 calculator and some increase in the program and computational effort on a large computer.

III. C 5 Two Missing Points Separated by a Single Point

When two missing points are separated by a single point, the obvious choices are between either smoothing first one missing point using the data segment centered on it and then the other missing point using the data segment centered on it and using the smoothed value for the other missing point, or smoothing both values simultaneously using the data segment centered on the point between them. This situation is illustrated in Table 8 and Figure 8. The results of smoothing first for the missing point at t = 2111 and, subsequently, for the missing point at t = 2113 are shown in Table 8. The results of smoothing first at t = 2113 and then at t = 2111 are also shown in Table 8. Smoothing both missing points simultaneously produced the results shown in Table 8 (last two columns). The results are summarized below.

Smoothing Procedure	Smoothed Values t = 2111	Smoothed Values t = 2113
Temp. Values	33,889.8	33,870.4
First at 2111	33,889.5	33,868.7
First at 2113	33,889.2	33,870.5
Simultaneous	33,887.9	33,870.3

Note that the greatest difference between the smoothed values is less than 2 units. If this difference is not considered to be serious then the simpler procedure of simultaneous smoothing could be preferred.

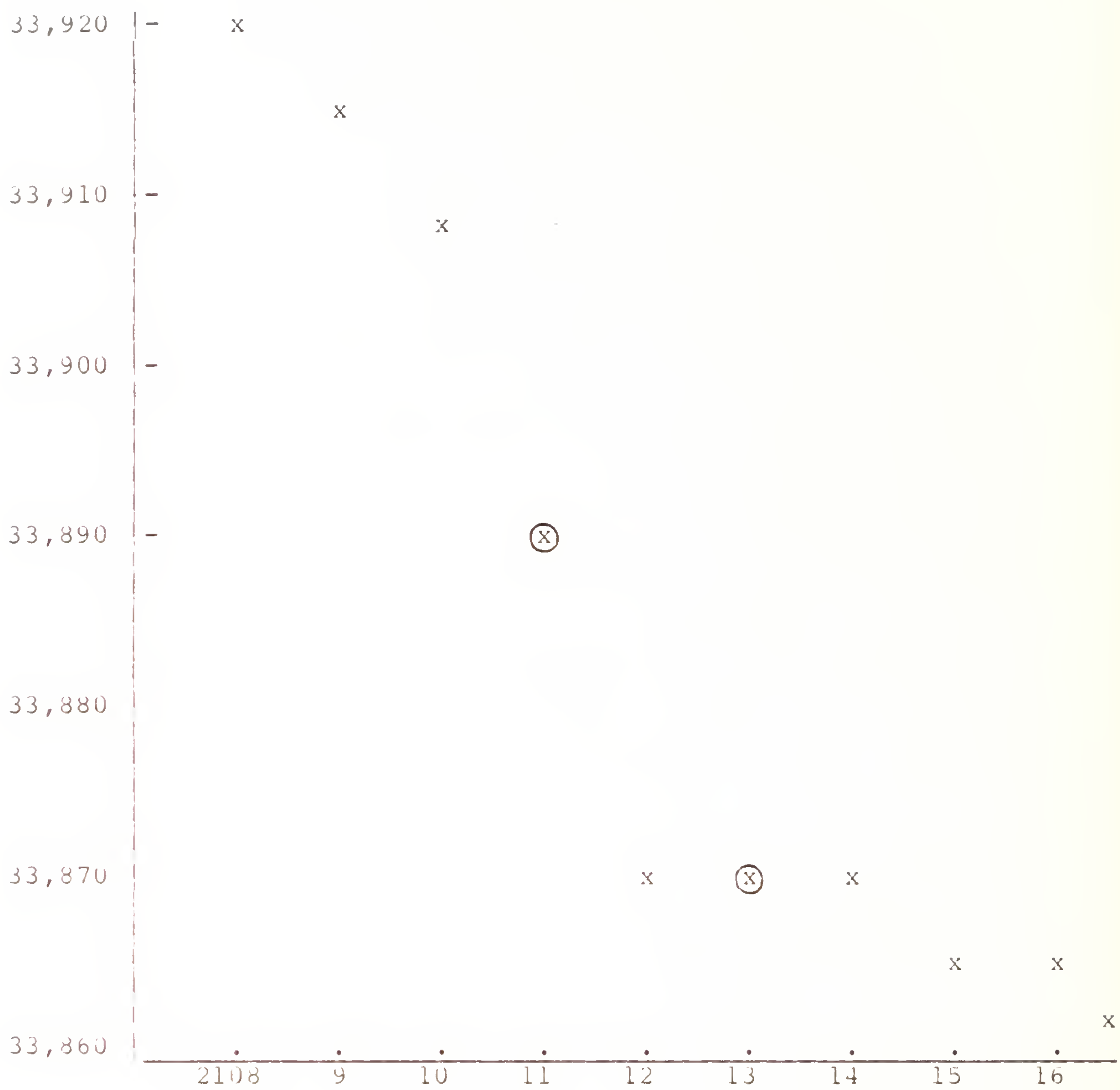


Figure 8. Separated Missing Points (2BX)

Table 8. Missing Points at t = 2111, 2113 (2BX)

t_i	x_i	Segment Center							
		2111	2113(1)	2113(2)	2113(1)	2113(2)	2111	2112(1)	2112(2)
2108	33,920.3								
2109	33,914.7								
2110	33,909.3								
2111 M	33,889.8	———	33,889.5	33,889.5			———		33.999.3
2112	33,870.3								
2113 M	33,870.4		———	33,869.1	———	33,869.1	33,868.7		33,870.6
2114	33,870.5								
2115	33,964.3								
2116	33,864.6								
SDR1		7.476	9.297	9.495	9.294	9.493	7.611	8.472	8.586
SDR2		7.780	5.151	5.096	5.143	5.091	7.917	6.075	6.017
SDR3		4.285	3.717	3.615	3.789	3.694	3.839	6.776	6.797
R		3	3	3	3	3	3	2	2
b_3		0.9306	-0.5472	-0.5472	-0.5389	-0.5389	0.9778	-----	-----
b_2		0.6667	1.9702	2.0321	1.9702	2.0321	0.6810	1.5857	1.6321
b_1		-16.4067	-2.7516	-2.7515	-2.8313	-2.8313	-16.8587	-8.8643	-8.8036
b_0		-2.6667	-7.8810	-8.1286	-7.8810	-8.1286	-2.7238	-6.3429	-6.5286
x'_{-3}		33,919.6	33,909.9	33,910.0	33,909.9	33,910.0	33,919.5	33,918.7	33,918.6
x'_{-2}		33,917.6	33,886.9	33,886.7	33,887.0	33,886.8	33,917.8	33,901.9	33,901.6
x'_{-1}		33,905.7	33,874.4	33,874.0	33,874.5	33,874.1	33,905.7	33,888.3	33,887.9
x'_{-0}		33,889.5	33,869.1	33,868.7	33,869.1	33,868.7	33,889.2	33,877.8	33,877.5
x'_1		33,874.7	33,867.8	33,867.4	33,867.7	33,867.4	33,874.0	33,870.6	33,870.3
x'_2		33,866.8	33,867.1	33,866.9	33,867.1	33,866.9	33,866.0	33,866.5	33,866.4
x'_3		33,871.4	33,863.8	33,863.9	33,863.3	33,864.0	33,871.1	33,865.5	33,865.8
r_{-3}		0.69	-0.57	-0.69	-0.62	-0.75	0.82	-4.01	-3.87
r_{-2}		-2.85	2.63	2.82	2.80	2.98	-3.10	7.39	7.69
r_{-1}		3.64	-4.074	-3.70	-4.19	-3.82	3.56	1.51	0.39
r_0		0.28	1.30	0.43	1.25	0.39	0.32	-7.54	-7.17
r_1		-4.41	2.72	3.10	2.75	3.12	-3.68	-0.16	0.30
r_2		3.58	-2.80	-2.62	-2.75	-2.57	2.70	4.04	4.11
r_3		-0.92	0.79	0.67	0.76	0.64	-0.63	-1.22	-1.45

There is some concern about this procedure, however, because of the large residual errors at t = 2112 and t = 2114. This concern is also supported by the large values of the SDR's (the Standard Deviations of the Residual Errors. On examination of Figure 8 it can be seen that there are two possible explanations of the large values of the SDR's. The first is that the actual vehicle track is inadequately represented by a cubic polynomial (Model error). The other is that the noise level in this path segment is greater than normal. The decision as to which explanation is correct cannot be determined from the data. Hopefully, vehicular control information and maneuver capabilities will be of use here.

Smoothing the value at t = 2112 should not be performed until after smoothed values have been established for the missing points so that its observed value is included in establishing their smoothed values. Only then should the value at t = 2112 be smoothed using the smoothed values for the missing points.

III. C 6 Missing Points Separated by Two Points

When missing points are separated by two observed values, simultaneous smoothing appears questionable since, whatever data segment is used, one of the missing points will not be adjacent to the center of the data segment. The preferred procedure would appear to be to smooth one of the missing points first using the data segment centered at that missing point, then do the same for the other point. If, when the second missing point is smoothed, the residual error of the smoothed value for the first missing point is large (arbitrarily, greater than unity) it would seem reasonable to resmooth the first again.

An example where two missing points are separated by two observed values occurs in the data for the second vehicle (2BX) where there are missing points at t = 2073 and t = 2076. The data and graph are presented in Table 9 and Figure 9. Smoothing first for the missing point at t = 2073 produced the results shown in Table 9. Since the residual error at t = 2076 is less than unity, the temporary value at t = 2076 was not subsequently smoothed using the smoothed value at t = 2073 as suggested above. Instead, the value at t = 2076 was smoothed using the temporary value for the missing point at t = 2073. Again, both residual errors were within the specified limit of unity. It would appear that the suggested procedure of smoothing first one missing point and then the other including the smoothed value for the first is not always necessary. That it was not in this example is no guarantee, however, that it may not be desirable in other cases.

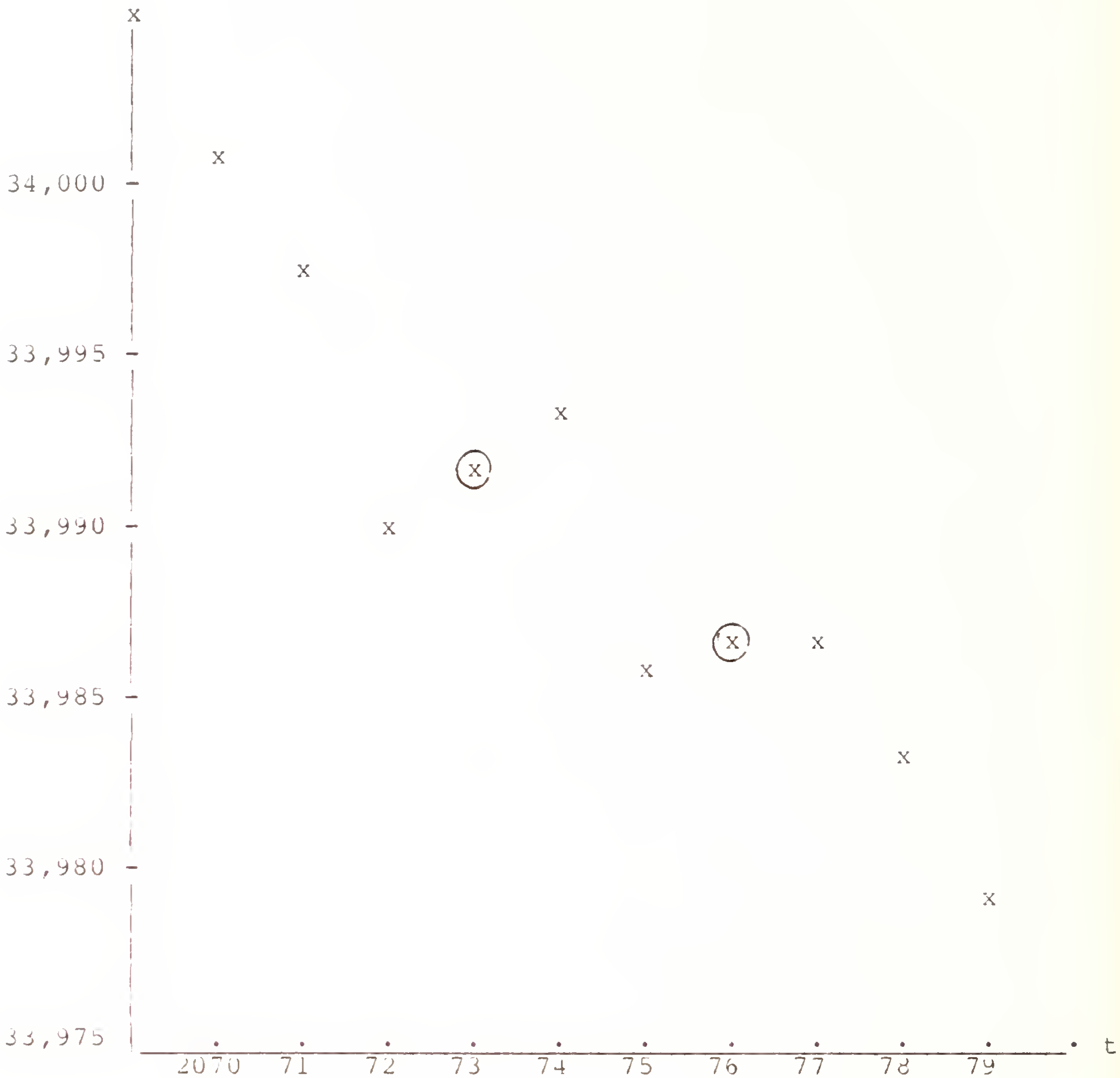


Figure 9. Separated Missing Points

Table 9. Missing Points Separated by Two Points (2BX)

t_i	x_i	Segment Center 2073	Segment Center 2076	Segment Center 2074	Segment Center 2075
2070	34,000.6				
2071	33,997.3				
2072	33,989.8				
2073 M	33,991.0				
2074	33,992.1				
2075	33,986.3				
2076 M	33,986.6				
2077	33,986.8				
2078	33,984.8				
2079	33,978.7				
SDR1		2.591	1.856	2.426	1.759
SDR2		2.517	1.681	2.478	1.914
SDR3		2.624	1.856	2.795	1.912
R		2	2	1	1
b_3		--	--	--	--
b_2		0.3131	-0.2655	--	--
b_1		-2.2036	-0.975	-1.524	-1.0321
b_0		-1.2524	-1.0619	33,990.0	33,988.2
x^1_{-3}		34,000.1	33,992.0	33,994.6	33,991.3
x^1_{-2}		33,996.4	33,989.7	33,993.0	33,990.3
x^1_{-1}		33,993.2	33,987.9	33,991.5	33,989.2
x^1_{0}		33,990.7	33,986.7	33,998.0	33,988.2
x^1_{1}		33,988.8	33,986.0	33,988.5	33,987.2
x^1_{2}		33,987.6	33,985.8	33,986.9	33,986.1
x^1_{3}		33,986.9	33,986.1	33,985.4	33,985.1
r_{-3}		0.47	-1.00	2.75	-1.50
r_{-2}		0.94	2.41	-3.23	0.74
r_{-1}		-3.42	-1.62	-0.51	2.87
r_{0}		0.30	-0.08	2.11	-1.9
r_{1}		3.29	0.83	-2.16	0.57
r_{2}		-1.25	-0.99	-0.39	0.66
r_{3}		-0.31	0.45	1.38	-0.30

As a further exploration of this example, simultaneous smoothing for the missing values at t = 2073 and t = 2076 was performed using data segments centered at t = 2074, and at t = 2075. The results are shown in Tables 9 also. In this example, data segments centered at any one of the four points appears to acceptable for establishing smoothed values for the missing points. Subsequent smoothing should, however, still be performed for the values at t = 2074 and 2075.

Although the SDR's are reasonably small for all choices of data segments, it is of some interest to compare the graph in Figure 9 with the one in the previous section (Figure 8). The scales on the y-axis are different but there appears to be some element of doubt about the actual path here also. Note that the smoothing procedure used second order polynomials to fit the segments centered at t = 2073 and 2076 but used first-order polynomials to fit the data segments centered at t = 2074 and 2075.

III D. TREATMENT OF MORE THAN TWO MISSING POINTS AND/OR POTENTIAL OUTLIERS.

1. General Discussion

The presence of more than three questionable values, either missing points or potential outliers, in a 7-point data segment cannot be smoothed to establish estimated values by a cubic equation. When there are three questionable values, they can be treated by either (1) iterated simultaneous smoothing or (2) establishing the cubic equation that fits the remaining four points in the segment exactly and then using that cubic equation be determine values for the three questionable points. When the same 7-point segment is used, the smoothing treatment (1) should converge to the exact fit (2). An example of this situation is explored in Section IIID2.

Similarly, if there are four questionable values in a given 7-point data segment the remaining three observations can be fitted exactly by a second order polynomial (parabola) or iterated simultaneous smoothing can be used to fit the parabola. Also, if there are five questionable values, the remaining two observations can be used to fit a first-order polynomial (a straight line) to these observations by either method.

It should be noted that the critical number of observations in a 7-point data segment required for fitting a polynomial of order k is $k+1$ since there are $k+1$ coefficients in the polynomial. If there are less than $k+1$ observations available then the polynomial cannot be established uniquely. If there are $k+1$ observations, it can either be fitted exactly or approximated by simultaneous iterated smoothing. If there are more than k+1 observations then only smoothing is appropriate.

It is also important to note that when a k^{th} order polynomial is fitted exactly to k+1 observations the standard deviation of the residual errors (SDR) is zero. In essence, the noise component is absorbed in the fitted polynomial and no estimate of the magnitude of the noise is possible. This absorption of the noise component into the target path is in contrast to situations (Section IIIB3 for example) where polynomials of order three or less provide inadequate representations of the vehicles path and hence part of the path variations are treated as noise. This results on larger standard deviations of the residual errors (SDR). It is worthy of emphasis, again, that a large value for SDR could be caused by either a large noise component or inadequacy of the polynomial model to represent the actual target path. It is important to determine which cause is pertinent. Potential sources for this information are internal control data for vehicular maneuvers, and examination of plots of the vehicle path. The latter would be difficult to incorporate into a data smoothing algorithm for automatic data processing (some human interaction may be necessary.) The use of internal control data appears to be a better approach of the goal of complete automation is to be achieved.

III D 2. THREE QUESTIONABLE VALUES

The problem with three questionable values in a 7-point data segment will be illustrated by the z-component of vehicle A where there are potential outliers at t = 2125 and t = 2127, and a missing value at t = 2128. A plot of values of the z_i's in a region containing possible 7-point data segments is shown in Figure 10 and listed in Table 10a. The fourth-order differences (Δ4) are listed in the third column.

Selection of the appropriate 7-point data segment is the first consideration. Centering it at t = 2125 would place the missing value at t = 2128 at the end of the segment and would not appear as desirable as centering it at t = 2126 or at t = 2127 to include the value on the other side of the missing point in the segment. Initially, it was decided to center the segment on the time between the potential outliers (t = 2126) so both potential outliers would be adjacent to the segment center and the missing point would not be an end point.

The 7-point L-S Polynomial Smoothing program was used to perform simultaneous iterative smoothing of the three questionable values with the results shown in Table 10a and the fourth column in Table 10a. Eight iterations were required to bring the residual errors of all three values within the prescribed level (r_i < 1.0). Fourth-order sequential differences were recalculated and are shown in column 5 of Table 10a. Note that no potential outliers are now indicated although the value of Δ4, at t = 2129 is close to the selected threshold of 50.

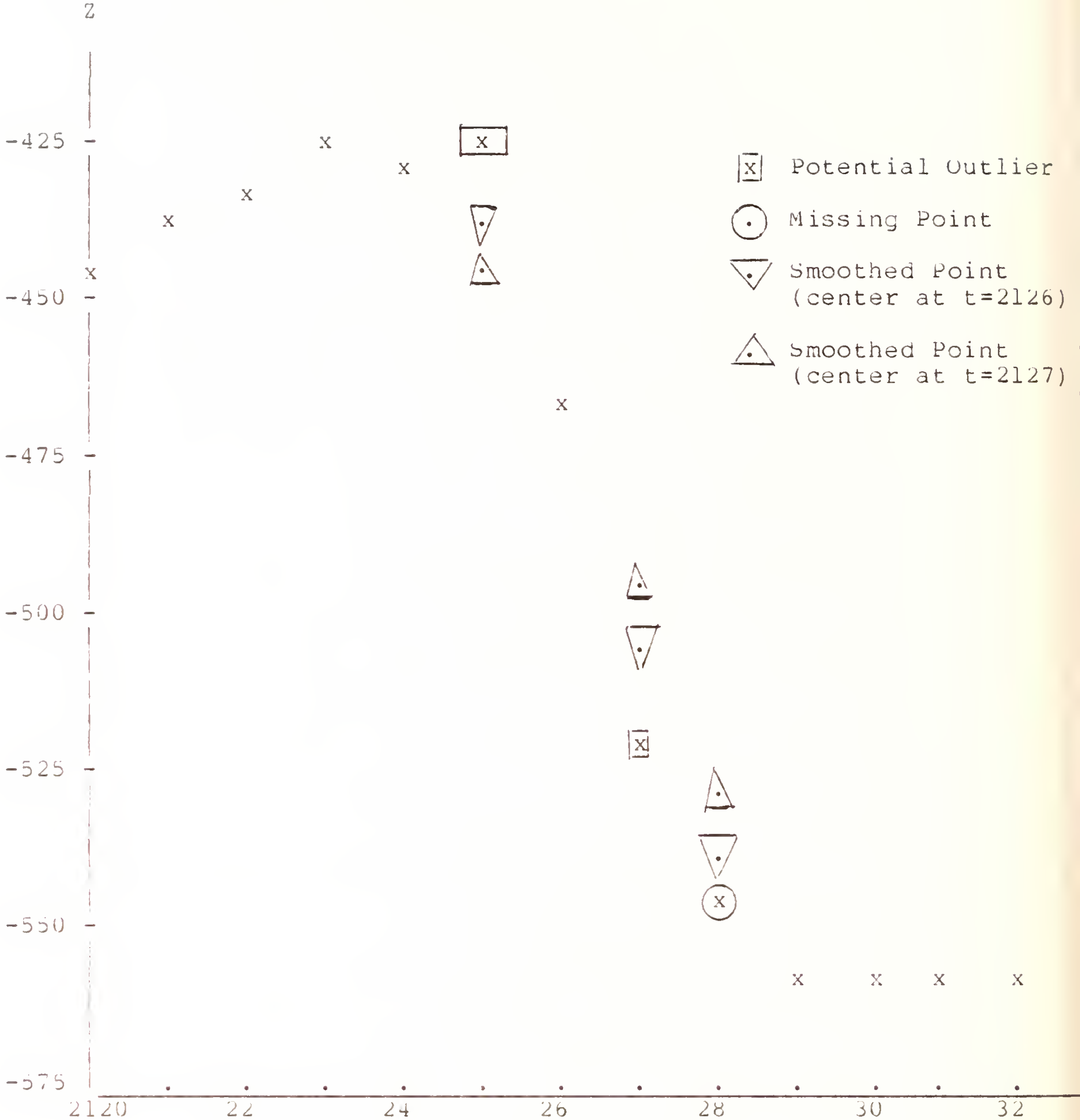


Figure 10. Three Questionable Values (2A Z)

Table 10a. 2A2 t = 2125 W, 2127 W, 2128M

t_i	Data		Smoothed Center at t = 2126		Exact Fit Center at t = 2126		Exact Fit Center at t = 2127	
	z_i	Δ4	$\hat{z}_i$	Δ4	z^*_i	Δ4	z^*_i	Δ4
2121	-440.1	+ 3.2						
2122	-434.7	-16.5		-16.5		-16.5		-16.5
2123	-426.3	+29.6		+12.4		+11.8		+10.4
2124	-428.3	-66.6		+ 2.2		+ 4.6		-13.8
2125	-424.5	+82.5 W	-441.7	- 1.2	-442.3	+ 0.1	-437.7	+20.8
2126	-465.3	+14.2		+11.6		- 0.1		- 0.2
2127	-518.6	-80.0 W	-498.7	- 8.0	-494.2	0.0	-501.1	+ 0.4
2128 M	-538.2	52.5	-530.7	+17.6	-526.0	+27.8	-535.3	- 0.4
2129	-557.7	-37.0		-46.9		-61.2 W		-30.9
2130	-558.5	+17.3		+24.7		+29.4		+20.1
2131	-558.8	+ 3.1		+ 3.1		+ 3.1		+ 3.1
2132	-559.6	-18.2						
2133	-558.8	+26.5						

Table 10b. SMOOTHING 2125 W, 2127 W, 2128 M (Center at t = 2126)

t_i	z_i	$z'_i(2)$	$z'(3)$	$z'(4)$	$z'(5)$	$z'(6)$	$z'(7)$	$z'(8)$
2123	-426.3							
2124	-428.3							
2125 W		-434.3	-437.8	-438.6	-440.5	-441.0	-441.3	-441.5
2126	-465.3	------	------	------	------	------	------	------
2127 W	-518.6	-509.0	-506.8	-505.0	-503.3	-501.9	-500.7	-499.6
2128 M	-538.2	-542.5	-539.9	-537.5	-535.6	-534.0	-532.7	-531.6
2129	-557.7							
2130	-558.5							
SDR 1	19.287	15.992	14.627	13.828	13.374	13.093	12.919	12.807
SDR 2	17.556	12.797	10.807	9.399	8.348	7.531	6.879	6.323
SDR 3	10.869	5.888	4.608	3.757	3.117	2.605	2.185	1.818
k	3	3	3	3	3	3	3	3
b_3	2.0167	1.5972	1.3667	1.2	1.075	0.9778	0.9	0.8333
b_2	-2.7321	-2.725	-2.6786	-2.6786	-2.7071	-2.7393	-2.7714	-2.8036
b_1	-39.4060	-36.0841	-34.0810	-32.6143	-31.5107	-30.6480	-29.9571	-29.3655
b_0	10.9286	10.9	10.7143	10.7143	10.8286	10.9571	11.0857	11.2143
z_{-3}	-429.7	-429.0	-428.4	-427.9	-427.6	-427.4	-427.2	-427.0
z_{-2}	-417.2	-421.1	-423.1	-424.3	-425.2	-425.7	-426.2	-426.6
z_{-1}	-434.3	-437.8	-439.6	-440.5	-441.0	-441.3	-441.5	-441.7
z_0	-468.9	-469.6	-469.6	-469.2	-468.7	-468.3	-467.8	-467.4
z_1	-509.0	-506.8	-505.0	-503.3	-501.9	-500.7	-499.6	-498.7
z_2	-542.5	-539.9	-537.5	-535.6	-534.0	-532.7	-531.6	-530.7
z_3	-557.3	-559.2	-559.0	-558.8	-558.6	-558.5	-558.3	-558.2
r_{-3}	3.44	2.68	2.05	1.61	1.30	1.07	0.89	0.74
r_{-2}	-11.14	-7.20	-5.23	-3.97	-3.15	-2.56	-2.11	-1.75
r_{-1}	9.76	3.52	1.75	0.91	0.51	0.33	0.23	0.17
r_0	3.61	4.29	4.29	3.94	3.44	2.96	2.51	2.1
r_1	-9.56	-2.20	-1.82	-1.66	-1.41	-1.23	-1.06	-0.86
r_2	4.32	-2.62	-2.37	-1.91	-1.61	-1.31	-1.09	-0.92
r_3	- .43	1.54	1.34	1.09	0.91	0.75	0.63	0.53

The four observations in this segment that were not considered questionable were next fitted by a cubic polynomial.

$$z^*(t_i') = b_0 + b_1 t' + b_2 t'^2 + b_3 t'^3$$

where

t_i	2123	2124	2125	2126	2127	2128	2129
t_i'	-3	-2	-1	0	+1	+2	+3

so that

t_i'	-3	-2	0	+3
z_i	-426.3	-428.3	-465.3	-557.7

The derivation of cubic equation fitting these four points exactly gives (Table 10c)

$$z^{*}(t') = -465.3 - 26.46t' - 2.96667t'^{2} + 0.506667t'^{3}.$$

Estimates to the values at the times of the questionable values ($t' = -1, +1, +2$) were established using this equation and are presented in column 6 of Table 10a. Sequential differences were recalculated and fourth order differences presented in column 7 of Table 10a.

Comparison of the values in columns 2, 4, and 6 indicate the following:

(a) The observed values at $t = 2125$ and $t = 2127$ are inconsistent with the rest of the observations (at $t' = 2125$, $z_i - \hat{z}_i = 17.3$ and $z_i - z_i^{*} = 17.9$, and at $t' = 2127$, $z_i - \hat{z}_i = 19.9$ and $z_i - z_i^{*} = 24.4$) so that both potential outliers should be reclassified as <u>actual</u> outliers.

(b) The smoothed values, $\hat{z}(t')$, are fairly close to the estimates $z^{*}(t')$ after eight iterations. More iterations should bring them still closer but the iterations were stopped when the residual errors were reduced to less than unity at all three suspect times.

The fourth order differences in column 7 of Table 10a indicate that there is a new potential outlier at time $t = 2129$. On reference to the graph (Fig. 10), it appears that the observation at this time is not necessarily an outlier but that there is a change in the path of the vehicle which cannot be adequately approximated by a cubic polynomial beyond this point.

Table 10c. Exact Solution: Segment center at t = 2126

t	t'	z_i	z_i^*
2123	-3	-426.3	-426.3
2124	-2	-428.3	-428.3
2125 W	-1	----	-442.3
2126	0	----	-465.3
2127 W	+1	----	-494.2
2128 M	+2	----	-526.0
2129	+3	-557.7	-557.7

Fit Cubic $z^*(t) = b_0 + b_1t' + b_2t'^2 + b_3t'^3$ $t' = -3,-2,0,+3$

(1) $z^*(-3) = b_0 - 3b_1 + 9b_2 - 27b_3 = -426.3$

(2) $z^*(-2) = b_0 - 2b_1 + 4b_2 - 8b_3 = -428.3$

(3) $z^*(0) = b_0 = -465.3$ $\underline{b_0 = 465.3}$

(4) $z^*(+3) = b_0 + 3b_1 + 9b_2 + 27b_3 = -557.7$

Substitute b_0(3) in (1), (2), (4)

(1') $b_1 - 3b_2 + 9b_3 = -13.0$

(2') $b_1 - 2b_2 + 4b_3 = -18.5$

(4') $b_1 + 3b_2 + 9b_3 = 030.8$

Solve (1') for b_1

(1") $b_1 = -13.0 + 3b_2 - 9b_3$

Substitute b_1''(1") in (2'), (4')

(2") $b_2 - 5b_3 = -5.5$

(4") $b_2 = -2.96667$ $\underline{b_2 = -2.96667}$

Substitute b_2 (4") in (2")

(2"') $b_3 = 0.506667$ $\underline{b_3 = 0.506667}$

Substitute b_2(4") and b_3(2"') in (1")

$\underline{b_1 = -26.46}$

As an exploratory exercise, the exact solution using the data segment centered at t = 2127 was also established. The results are presented in columns 8 and 9 in Table 10a. It is interesting to note that the observed value at t = 2129 does not appear as a potential outlier in the recalculated fourth order sequential differences. Neither does the observation at t = 2130. Since subsequent fourth order differences are not affected, there is no potential outlier remaining when this data segment is used.

IV. Conclusions and Recommendations

Questionable data values, either potential outliers or temporary values for missing points, degrade the quality of smoothed estimates of points on a vehicular path. A position location system which omits observations at every eighth observational time (scheduled missing points) makes the treatment of other questionable values more difficult and, if the latter are frequent, can even preclude the use of smoothing.

Although potential outliers are treated the same way as missing points in smoothing, a specific data segment, they can produce greater contamination of the smoothing process and should be given priority in any smoothing algorithm. Also, on replacement of a potential outlier by a smoothed value, sequential differences should be recalculated to determine whether other potential outliers occur in its vicinity. It is important, wherever possible, to establish whether a potential outlier is actually an outlier (a wild observational value) or is an indicator of a change in a vehicular path that cannot be adequately represented by a polynomial of order three or less. Automation of this identification of the cause for a potential outlier may be facilitated by other sources of information on changes in vehicular paths such as internal control data. An alternative source of this information is manual observation of a plot of the observed data points to establish points at which the vehicular path has changed so that it cannot be expected to be represented by a polynomial of order three or less. (The latter reduces the extent to which automation can be achieved and hence the incorporation of internal control data into the smoothing process is preferred.)

Isolated questionable values cause little problem since they can be treated simply by iterated smoothing to establish replacement estimated values consistent with the other observations in the 7-point data segment centered at the time of the questionable value. The presence of more than one questionable value requires more complex treatment. Occurrence of two or three such values require different treatments and was discussed separately. If more than three questionable values occur in a 7-point data segment the 7-point least squares smoothing procedure is not applicable. (Polynomials of order one or two could still be considered depending on the number of questionable values but should be avoided since their ability to represent actual vehicular paths is questionable.) Such data segments should be identified for both potential users of the smoothed data and data collectors.

When there are two questionable values close to each other both nature, missing point or potential outlier, and their time separation need to be considered in establishing the appropriate treatment. The following cases and their treatments appear reasonable:

a. Adjacent questionable values.

(1) If the two questionable values consist of a potential outlier and a missing point, then the two should be smoothed simultaneously using the data segment centered at the time of the potential outlier,

(2) If two adjacent questionable values are both missing points, then they should also be smoothed simultaneously using the data segment centered at the time of one of them. (The choice of center may affect the resulting smoothed values but no general rule for preference can be given.)

(3) Situations in which adjacent questionable values are both potential outliers appears to be unlikely so it is not considered.)

b. Two questionable values separated by a single observation.

For the reason of simplicity of the smoothing algorithm and reduction in computation the two values should be smoothed simultaneously using the data segment centered at the observation time between the two questionable values.

c. Two questionable values separated by more than one observation.

Since at least one of the questionable values cannot be adjacent to the 7-point segment center, it would be reasonable to smooth first one, then the other, returning to the first for resmoothing. Priority of smoothing is for potential outliers and, if both are potential outliers, the first smoothed should be the one with the largest fourth order sequential difference.

Situations involving three questionable values could be smoothed simultaneously using a data segment centered so that all three are as close to the center of the segment as possible. A substantial number of iterations may be required to bring the three residual errors to within the specified level. It would appear preferrable here to omit smoothing and to fit the remaining four points in the data segment using simultaneous linear equations to determine the coefficients of the cubic equation to fit these four points exactly. (It would be possible to use smoothing limiting the polynomial to order two or less but, again, the question of adequate representation of the target path arises.) Whether simultaneous smoothing or the exact fit is used, the procedure, in essence, treats the noise components of the four observations as part of the vehicular path instead of noise. Thus a reduction in the quality of the estimates is introduced and this information should be indicated to both potential users and data collectors.

The material presented in this report has emphasized details which should be useful in understanding the smoothing process and in implementing an appropriate program for smoothing 3-D data at NUWES. It also provides essential background for an investigation of the quality of 3-D data and for the establishment of Figures of Merit for 3-D data submitted for smoothing which is to follow.

References

1. J. B. Tysver, "Smoothing 3-D data for torpedo paths", Naval Postgraduate School Technical Report NPS55-78-036Pr, May 1978.

2. J. B. Tysver, "Use of sequential differences in smoothing 3-D data", Naval Postgraduate School Technical Report NPS55-79-012Pr, May 1979.

3. J. B. Tysver, "A 3-D Data Smoothing Algorithm", Naval Postgraduate School Technical Report NPS55-81-001, January 1981.

4. J. B. Tysver, "Selection of Segment Length for Least-Squares Polynomial Smoothing", Naval Postgraduate School Letter Report, January 1982.

DISTRIBUTION LIST

	No. of Copies
Commanding Officer Naval Undersea Warfare Engineering Station Keyport, WA 98345	
Attn: Mr. R. L. Marimon, Code 70	1
CDR Brian Uber, Code 80	1
Mr. R. L. Mash, Code 50	1
Mr. S. McKeel, Code 51	1
Mr. Pal R. Hughes, Code 5122	1
Library, Code 0142 Naval Postgraduate School Monterey, CA 93943	4
Dean of Research Code 012A Naval Postgraduate School Monterey, CA 93943	1
Library, Code 55 Naval Postgraduate School Monterey, CA 93943	1
Professor J. B. Tysver Code 55TY Naval Postgraduate School Monterey, CA 93943	3